CALIBRATION OF PROBABILITIES:

THE STATE OF THE ART

OREGON RESEARCH INSTITUTE
BRUNEL UNIVERSITY

Sarah Lichtenstein
Baruch Fischhoff
L.D. Phillips

ADVANCED DECISION TECHNOLOGY PROGRAM

CYBERNETICS TECHNOLOGY OFFICE
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
Office of Naval Research • Engineering Psychology Programs

The objective of the Advanced Decision Technology Program is to develop and transfer to users in the Department of Defense advanced management technologies for decision making. These technologies are based upon research in the areas of decision analysis, the behavioral sciences and interactive computer graphics. The program is sponsored by the Cybernetics Technology Office of the Defense Advanced Research Projects Agency and technical progress is monitored by the Office of Naval Research – Engineering Psychology Programs. Participants in the program are:

Decisions and Designs, Incorporated
The Oregon Research Institute
Perceptronics, Incorporated
Stanford University
The University of Southern California

Inquiries and comments with regard to this report should be addressed to:

Dr. Martin A. Tolcott
Director, Engineering Psychology Programs
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

or

LT COL Roy M. Gulick, USMC
Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

TECHNICAL REPORT DDI-3

CALIBRATION OF PROBABILITIES:

THE STATE OF THE ART

by

Sarah Lichtenstein, Baruch Fischhoff
Oregon Research Institute

and

L.D. Phillips
Brunel University

Sponsored by

Defense Advanced Research Projects Agency
Contract N00014-76-C-0074
ARPA Order No. 3052
Under Subcontract from
Decisions and Designs, Incorporated

August, 1976

OREGON RESEARCH INSTITUTE
P.O. Box 3196
Eugene, Oregon 97403
(503) 484-2123

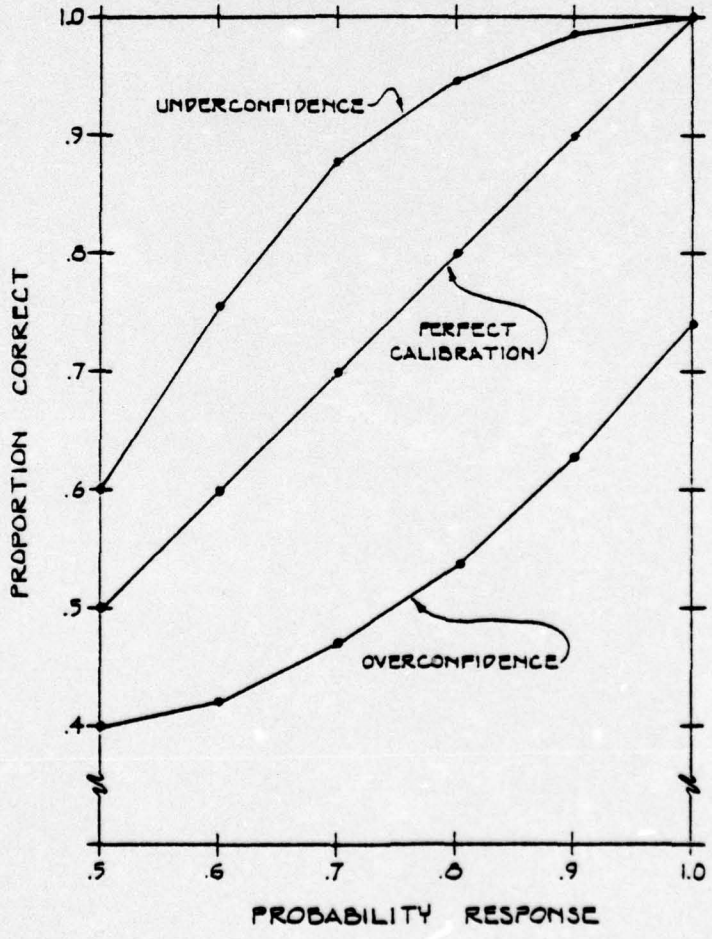
SUMMARY

Introduction

This paper presents a comprehensive review of the research literature on an aspect of probability assessment called "calibration." Calibration measures the validity of probability assessments. Being well-calibrated is critical for optimal decision-making and for the development of decision aiding techniques.

Background and Approach

Subjective probability assessments play a key role in decision making. It is often necessary to rely on an expert to assess the probability of some future event. How good are such assessments? One important aspect of their quality is called calibration. Formally, an assessor is calibrated if, over the long run, for all statements assigned a given probability (e.g., the probability is .65 that "Romania will maintain its current relation with People's China."), the proportion that is true is equal to the probability assigned. For example, if you are well calibrated, then across all the many occasions that you assign a probability of .8, in the long run 80% of them should turn out to be true. If, instead, only 70% are true, you are not well calibrated, you are overconfident. If 95% of them are true, you are underconfident. The figure below shows calibration curves of well-calibrated, overconfident and underconfident assessors.



While this characteristic of assessors has obvious importance for applied situations, people's calibration has rarely been discussed by decision analysts or decision advisors. In the last few years, there has developed an extensive literature about calibration, reporting both laboratory and real-world experiments. It is now time to review this literature, to look for common findings which can be used to improve decisions, and to identify unsolved problems.

Findings

Two general classes of calibration problem have been studied. The first class is calibration for events for which the outcome is discrete. These include probabilities assigned to statements like "I know the answer to that question," "They are planning an attack," or "Our alarm system is foolproof." For such tasks, the following generalizations are justified by the research:

1. Weather forecasters, who typically have had several years of experience in assessing probabilities, are quite well calibrated.

2. Other experiments, using a wide variety of tasks and subjects, show that people are generally quite poorly calibrated. In particular, people act as though they can make much finer distinctions in their degree of uncertainty than is actually the case.

3. Overconfidence is found in most tasks; that is, people tend to overestimate how much they know.

4. Despite the abundant evidence that untutored assessors are badly calibrated, there is little research showing how and how well these deficiencies can be overcome through training.

The second class of tasks is calibration for probabilities assigned to uncertain continuous quantities. For example, what is the mean time between failures for this system? How much will this project cost? The assessor must report a probability density function across the possible values of such uncertain quantities. The usual method for eliciting such probability density functions is to assess a small number of fractiles of the function. The .25 fractile, for example, is that value of the uncertain quantity such that there is just a 25% chance that the true value will be smaller than the specified value. Suppose we had a person assess a large numberof .25 fractiles. He would be giving numbers such that, for example, "There is a 25% chance that this repair will be done in less than x_1 hours" or "There is a 25% chance that Warsaw Pact personnel in Czechoslovakia number less than x_1." This person will be well calibrated if, over a large set of such estimates, the true value will be less than x_1 25% of the time. The measures of calibration used most frequently in research consider pairs of extreme fractiles. For example, experimenters assess calibration by asking whether 98% of the true values fall between an assessor's .01 and .99 fractiles.

For calibration of continuous quantities, the following results summarize the research.

1. A nearly universal bias is found: assessors' probability density functions are too narrow. For example, 20 to 50% of the true values lie outside the .01 and .99 fractiles, instead of the prescribed 2%. This bias reflects overconfidence; the assessors think they know more about the uncertain quantities than they actually do know.

2. Some data from weather forecasters suggests that they are not overconfident in this task. But it is unclear whether this is due to training, experience, special instructions, or the specific uncertain quantities they deal with (e.g., tomorrow's high temperature).

3. A few studies have indicated that, with practice, people can learn to become somewhat better calibrated.

Implications

Since assessed probabilities are central to a wide variety of decision problems (e.g., making intelligence estimates, assessing system reliability, projecting costs, deciding whether to acquire more information), the question of whether such probabilities are calibrated has far-reaching importance. Almost all decision analyses involve probability assessments. If these assessments are in error, the finest analysis relying on them may be faulty. The bias towards overconfidence reported here is widespread and well documented. What is not so well established is whether, and how, this bias can be overcome through training. The superior performance of weather forecasters is encouraging. These people have been using probabilities in their forecasts on a daily basis for several years; one might assume that this experience accounts for their excellence. Further research is needed to document just how much training, with what kind of feedback, is most efficient for improving assessors' calibration. Such research is crucial to developing a viable decision analysis technology. It also helps tell us how much faith to put in the probability assessments and decisions of untrained decision makers working without the benefit of decision aids.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENT

Support for this research performed by Oregon Research Institute was provided by the Advanced Research Projects Agency of the Department of Defense and was monitored under Contract N00014-76-C-0074 with the Office of Naval Research, under subcontract from Decisions and Designs, Inc.

CALIBRATION OF PROBABILITIES: THE STATE OF THE ART

INTRODUCTION

From the subjective point of view (de Finetti, 1937) a probability is a degree of belief in a proposition whose truth has not been ascertained. A probability expresses a purely internal state; there is no "right" or "correct" probability that resides somewhere "in reality" against which it can be compared. However, in many circumstances, it may become possible to verify the truth or falsity of the proposition to which a probability was attached. Today, we assess the probability of the proposition "it will rain tomorrow." Tomorrow, we go outside and look at the rain gauge to see whether or not it has rained. When verification is possible, we can use it to gauge the adequacy of our probability assessments.

Assessors' adequacy has been discussed by Winkler and Murphy (1968a), who identified two general kinds of "goodness," normative goodness, which reflects the degree to which the assessments conform to the axioms of probability and espress the assessor's true beliefs, and substantive goodness, which reflects the amount of knowledge of the topic area contained in the assessments. This paper reviews the literature about the kind of adequacy called calibration.

If a person assesses the probability of a proposition's being true as .7, and later finds that the proposition is false, that in itself does not invalidate the assessment. However, if a judge assigns .7 to 10,000 independent propositions, only 25 of which subsequently are found to be true, there is something wrong with these assessments. The attribute which they lack we call calibration. This attribute has also been called "realism" (Brown and Shuford, 1973), "realism of confidence" (Adams and Adams, 1961), "appropriateness of confidence" (Oskamp, 1962), "external validity" (Brown and Shuford, 1973), "secondary validity" (Murphy and Winkler, 1971), and "reliability" (Murphy, 1973). Formally, a judge is calibrated if, over the long run, for all propositions assigned a

given probability, the proportion that is true is equal to the probability assigned. We can empirically evaluate judges' calibration by observing their probability assessments, verifying the associated propositions and then observing the proportion that is true in each response category. Judges who are not calibrated may be either underconfident or overconfident. For the underconfident assessor, the proportion of propositions that are true is greater than the probability assigned to them. With overconfidence, too few propositions are true.

In this paper, we review the experimental literature on calibration, separated somewhat arbitrarily into two sections. The first is devoted to the calibration of assessors making probability judgments about discrete propositions; the second, to calibration for probability density function concerning uncertain numerical quantities. The arbitrariness arises from the fact that an uncertain quantity, for example, "the population of Brazil," can always be reworded into one or more discrete propositons, such as "the population of Brazil exceeds 85 million." In a few cases, our decision about which section an experiment should be discussed in depended more on how the authors reported their data than on how their subjects perceived the task.

Calibration is essentially a property of single individuals. Most of the results reviewed here, however, are grouped across subjects. Although grouping is often necessary to secure the large quantities of data needed for stable estimates of calibration, it can both obscure interesting individual differences and cause serious biases in studies in which only a few items are presented to many subjects. The experiemnter who relies on but a few stimuli may run the risk of inadvertently including a preponderance of items which most subjects answer incorrectly (e.g., Are potatoes native to Ireland or Bolivia? How many people live in Outer Mongolia?). With such "deceptive" items, perfect calibration is impossible. A large number of items is one

protection against this problem.

DISCRETE PROPOSITIONS

Discrete propositions can be stated with any number of alternatives:

No alternatives: What is absinthe? The subject is asked to provide an answer, and then to give the probability that his or her answer is correct. The entire range of probability responses, from 0 to 1, is appropriate. Only Adams (1957) has looked at calibration for this task.

One alternative: Absinthe is a precious stone. What is the probability that this statement is true? Again, the relevant range of the probability scale is 0 to 1.

Two alternatives: Absinthe is (a) a precious stone; (b) a liqueur. With the "half-range" method, the subject first selects the more likley alternative, and then states the probability that this choice is correct. This response must be $\geq .5$. With the "full-range" method, the subject gives the probability that a prespecified alternative is correct. Here the subject may use any response from 0 to 1.

Three or more alternatives: Absinthe is (a) a precious stone; (b) a liqueur; (c) a Caribbean island; (d) . . . Two variations of this task may be used: (1) the subject selects the single most likely alternative and states the probability that it is correct, using a response $\geq 1/k$ for k alternatives; (2) the subject assigns probabilities to all alternatives, using the range 0 to 1. This procedure induces dependencies in the data, by requiring the k assessments to sum to 1.

For all these variations, calibration may be reported via a "calibration curve." Such a curve is derived as follows: (1) Collect many probabilistic responses to items whose correct answer is known or will shortly be known to the experimenters. (2) Categorize the responses, usually within ranges; for

example, all responses between .60 and .69 are placed in the same category. (3) Compute for each category the proportion correct, that is, the proportion of items for which the proposition is true. (4) For each category, plot the mean response against the proportion correct.

Several measures of overall calibration have been proposed. Murphy (1973) has looked at the general case of k-alternative items. Each response, i, is represented by a row vector of probabilities, $\underline{r}_i = (r_1, \ldots, r_k)$, and the associated outcome by a row vector $\underline{c}_i = (c_{1i}, \ldots, c_{ji}, \ldots, c_{ki})$, where c_{ji} equals one for the true alternative and zero otherwise. Given response vectors for N items form a single individual, the Brier (1950) scoring rule (proper quadratic scoring rule such that the smaller the score, the better) is:

$$B = \frac{1}{N} \sum_{i=1}^{N} (\underline{r}_i - \underline{c}_i)(\underline{r}_i - \underline{c}_i)' \quad ,$$

in which the prime denotes a column vector. Murphy partitioned this score into three terms. The response vectors are sorted into T subcollections such that all the responses $\underline{r}_t$ in a subcollection are identical. Let n_t be the number of responses in the t'th subcollection, and let $\bar{\underline{c}}_t$ be the proportion-correct vector for the t'th subcollection:

$$\bar{\underline{c}}_t = (\bar{c}_{1t}, \ldots, \bar{c}_{jt}, \ldots, \bar{c}_{kt}) \quad , \text{ where } \quad \bar{c}_{jt} = \sum_{t=1}^{n_t} c_{jt}/n_t \quad .$$

Let $\bar{\underline{c}}$ be the proportion-correct vector across all responses,

$$\bar{\underline{c}} = (\bar{c}_1, \ldots, \bar{c}_j, \ldots, \bar{c}_k) \quad , \text{ where } \quad \bar{c}_j = \frac{1}{N} \sum_{i=1}^{N} c_{ji} \quad ,$$

and let $\underline{u}$ be the unity vector, a row vector whose k elements are all one.

Then Murphy's partition of the Brier score is:

$$B = \bar{\underline{c}}(\underline{u} - \bar{\underline{c}})' + \frac{1}{N} \sum_{t=1}^{T} n_t (\underline{r}_t - \bar{\underline{c}}_t)(\underline{r}_t - \bar{\underline{c}}_t)' - \frac{1}{N} \sum_{t=1}^{T} n_t (\bar{\underline{c}}_t - \bar{\underline{c}})(\bar{\underline{c}}_t - \bar{\underline{c}})'$$

The first term measures the uncertainty inherent in the set of N items. For example, if all items concern rain vs. no rain, this term reflects how often it rained in fact. The second term, which Murphy called "reliability," is a measure of calibration, the weighted sum of squares of the difference between the responses and the proportion correct for those responses. The third term, called "resolution," reflects the ability of the assessor to sort the events into subcategories for which the hit rate is maximally different from the overall hit rate.

Murphy (1974) has further suggested a "sample skill score" to measure the skill of forecasters. This score, which constitutes a proper scoring rule, is calculated by subtracting the second term in the partition, calibration, from the third term, resolution. Assessors should maximize this score; the maximum is (k-1)/k.

Murphy's partition was designed for repeated predictions of the same event, e.g., rain. When the items are diverse, as in a multiple-choice examination, so that the alternatives can be identified only as "first alternative, second alternative," and so forth, then the first term is not meaningful; it is simply a function of the order in which the true alternatives were arranged across items.

When the assessor is asked first which is the correct alternative, and next what the probability is that the chosen alternative is correct, only one response per item is scored. In these cases, Murphy's (1974) measure reduces to what he has called (Murphy, 1972) the "special scalar partition:"

$$B' = \bar{c}(1-\bar{c}) + \frac{1}{N}\sum_{t=1}^{T} n_t(r_t-\bar{c}_t)^2 - \frac{1}{N}\sum_{t=1}^{T} n_t(\bar{c}_t-\bar{c})^2 \quad ,$$

where $\bar{c}$ is the overall proportion correct, and $\bar{c}_t$ is the proportion correct in the t'th subcategory. When the second response is the response $\geq .5$ (as

with the two-alternative, half-range task), the first term does have an interpretation: it reflects the subject's ability to pick the correct alternative, and thus might be called "knowledge." The second term measures calibration, and the third, resolution, as before.

This scalar measure of calibration, a weighted squared error, is similar to measures proposed by Adams and Adams (1961), who used a "mean absolute discrepancy score,"

$$\sum_{t=1}^{T} \sqrt{n_t}\ |r_t - \bar{c}_t| \ / \sum_{t=1}^{T} \sqrt{n_t} \quad ,$$

and by Oskamp (1962), who used an "appropriateness of confidence" scale:

$$\frac{1}{N} \sum_{t=1}^{T} n_t\ |r_t - \bar{c}_t| \quad .$$

Shuford and Brown (1975) also started with a proper scoring rule, the logarithmic. In addition to computing a score for the assessor's responses, S, they proposed fitting a least squares regression line to the data in a calibration curve. The equation for the best-fitting line can be used to externally recalibrate the assessor's responses, in order to correct for systematic bias. One can then compute the score for these recalibrated responses, $\hat{S}$. If M is the maximum score possible, then $M-\hat{S}$ measures the loss in score due to lack of knowledge, while $\hat{S}-S$ measures the loss in score due to poor calibration.

None of these measures of calibration have as yet gained acceptance in the research literature. None discriminate overconfidence from underconfidence. Nothing is known about the sampling properties of any of the measures.

Meteorological Research

In 1906, W. Ernest Cooke, Government Astronomer for Western Australia, advocated that each meteorological prediction be accompanied by a single number which would "indicate, approximately, the weight or degree of

probability which the forecaster himself attaches to that particular prediction." He reported (Cooke, 1906a, 1906b) results from 1,951 predictions. Of those to which he had attached a weight of 5 ("almost certain to be verified"), .985 were correct. For his weight of 4 ("normal probability"), .938 were correct, while for his weight of 3 ("doubtful"), .787 were correct.

In 1951, Williams asked eight professional Weather Bureau forecasters in Salt Lake City to associate the number 0, .2, .4, .6, .8, or 1.0 with each 12-hour forecast of precipitation. The calibration curve for 1,095 predictions appears in Figure 1. These assessments of the probability of precipitation were too high throughout most of the range (see Figure 1). This might be the result of a fairly natural form of hedging in public pronouncements. People are much less likely to criticize a weather forecast that leads them to carry an umbrella when it does not rain than one that leads them to be without an umbrella when it does.

Similar results emerged from two studies of forecasters reported by Murphy and Winkler (1974). One of their studies dealt with the effect of a computerized weather prediction system (PEATMOS) on forecasters' assessments. The task was to assess the probability of precipitation the following day. Forecasters did this twice, before and then again after seeing the PEATMOS output. Data were collected in Great Falls, Montana, and Seattle, Washington. All 7,188 assessments (before and after PEATMOS in both cities) were combined to produce the calibration curve in Figure 1, which shows the same overestimation of the probability of rain.

In the other study forecasters were asked to predict the next day's high temperature. Two forecasters used a "fixed-width, variable-probability" technique. First, they named the median temperature. Then they stated the probability that the temperature would fall within intervals of 5° F and 9° F centered at the median. Such a technique converts a continuous probability

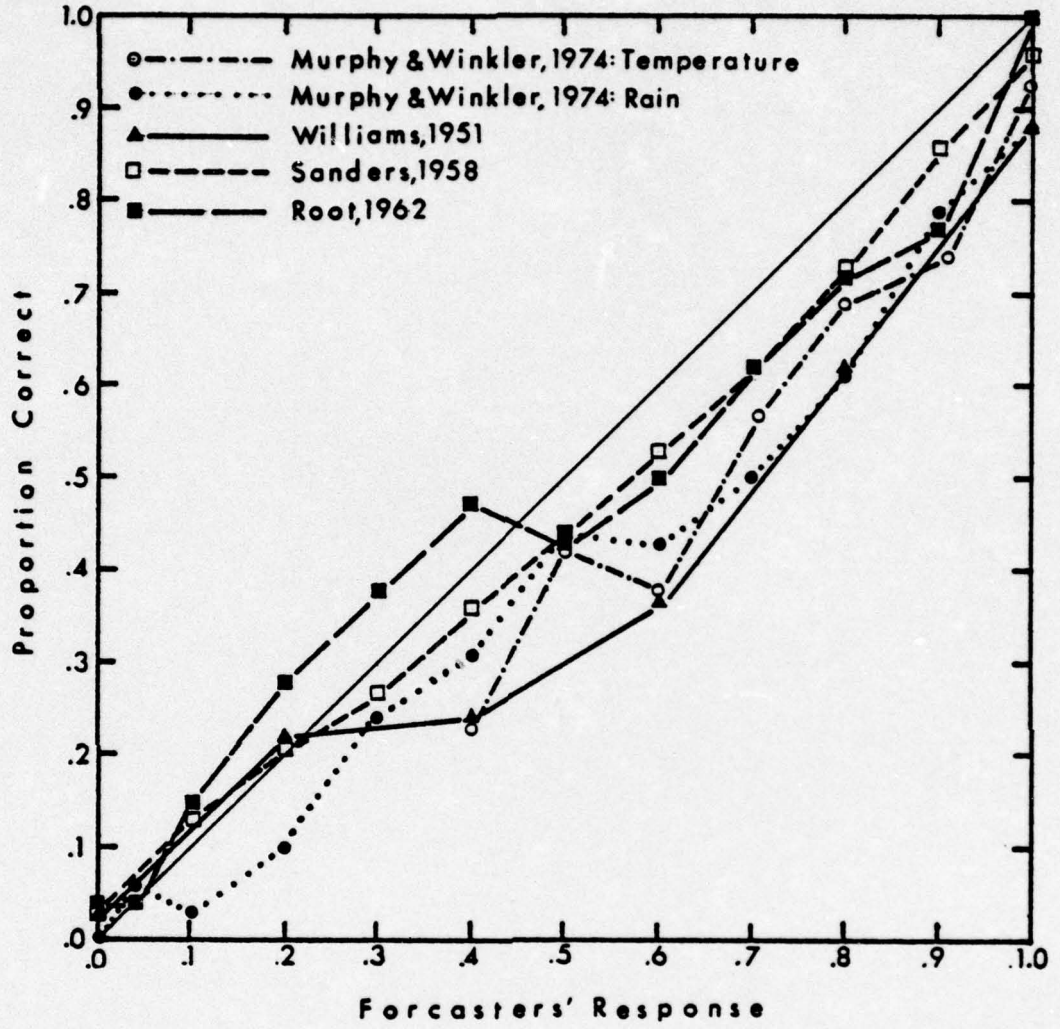


Figure 1

Calibration of Weather Forecasters

distribution into a two-alternative discrete task: the temperature is scored as falling either within or outside of the stated interval. Calibration for 241 such assessments is shown in Figure 1. These forecasters, who could have used any probability between 0 and 1, responded below .4 on only three occasions (excluded from the curve). Again, we see a systematic bias across the entire range covered: the probability associated with the temperature falling inside the interval is always too large. Better calibration was reported by Sanders (1958) who collected 12,635 predictions of a variety of dichotomized events: wind direction, wind speed, gusts, temperatures, cloud amount, ceiling, visibility, precipitation occurrence, precipitation type, and thunder-storm, using the eleven responses 0, .1, ...9, 1.0. The resulting calibration curve is shown in Figure 1.[1]

In contrast to the meteorological studies showing a constant bias across almost the entire response range, Root (1962) has reported calibration for 4,138 precipitation forecasts which shows (see Figure 1) a more systematic pattern. Here, assessed probabilities were too low in the low range and too high in the high range, relative to the observed frequencies. This pattern indicates overconfidence both for the proposition, "It most likely will rain," and for the proposition, "It most likely won't rain."

Figure 2 shows calibration curves for one year of precipitation probability forecasts from Hartford, Connecticut (Winkler and Murphy, 1968b). These forecasters had the option of forecasting for either a six-hour period or a twelve-hour period. They made 3,174 six-hour forecasts and 2,936 twelve-hour forecasts; these data are shown separately. There was some ambiguity about whether the forecasters had intended to include or exclude

[1] The references by Cooke (1906), Williams (1951), and Sanders (1958) were brought to our attention through an unpublished manuscript by Howard Raiffa, dated January, 1969, entitled "Assessments of probabilities."

"a trace of precipitation" (less than .01 inches) in their predictions of precipitation. Accordingly, the data were analyzed twice, once assuming that "precipitation" included the occurrence of traces, and once assuming that "precipitation" did not include traces. The inclusion or exclusion of traces had a substantial effect, as did the choice of time period. Six-hour forecasts were associated with lower observed frequencies than were twelve-hour forecasts. Thus the forecasters were found to assess precipitation probabilities that were too high for the six-hour, traces excluded case, and too low for the twelve-hour, traces included case, relative to the observed frequencies, while the other two cases showed very good calibration.

The United States Weather Bureau (1969) has collected massive amounts of calibration data for precipitation forecasts made from April, 1967, to March, 1968, at sites all over the country. Figure 3 shows just one-fourth of these data (the rest of the data were highly similar); each curve is based on more than 16,800 forecasts. The solid-line curve is for forecasts for the first time period, that which immediately followed the time the forecast was made. Here, calibration was excellent, with a mean absolute error of only .03. As the lag between the time the forecast was made and the period it referred to increased, calibration deteriorated. This deterioration was not as great as it appears in the figure, because in the later periods forecasters used fewer responses in the high range. Thus, even for the third period the mean absolute error was only .05. Murphy[2] believes that these data more accurately represent the current performance of weather forecasters than do the data in Figures 1 and 2. He attributes the superior performance in the present report to the increased experience with probabilities that the forecasters have gained over the years, and to the fact that these data were gathered from real on-the-job performance, whereas some of the previous data either were collected in experimental situations (Winkler and Murphy, 1974) or with events that are not usually forecasted probabilistically (Sanders, 1958).

[2]Personal communication, February, 1976.

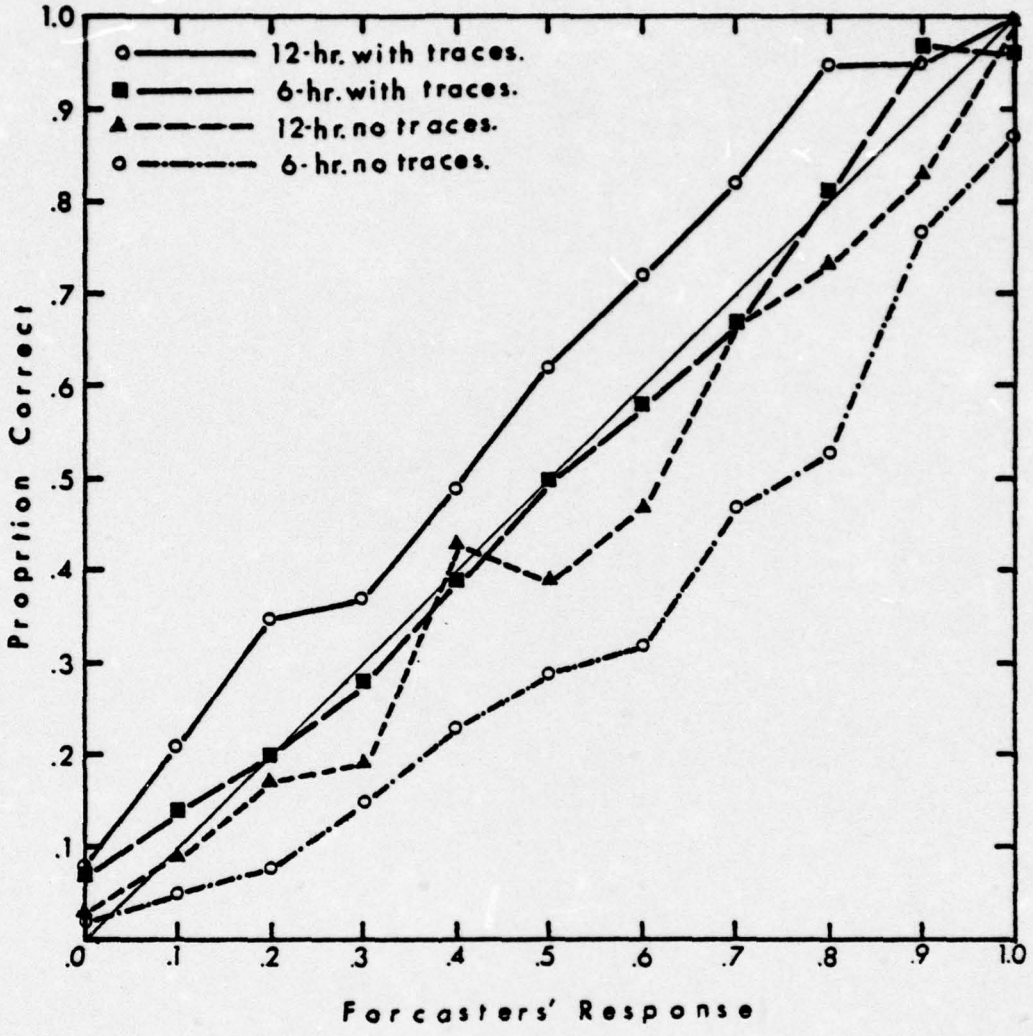


Figure 2

Calibration for Precipitation Forecasts

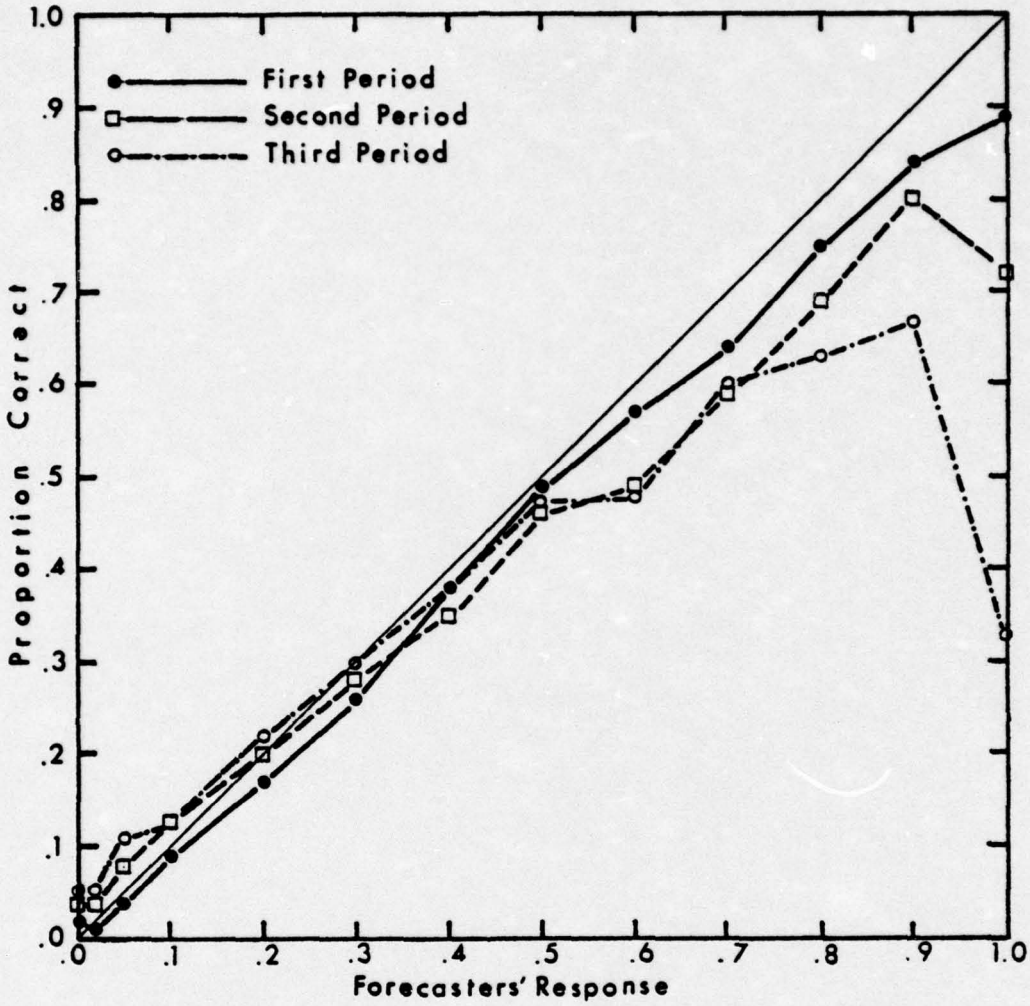


Figure 3

U.S. Weather Bureau Calibration as a Function of Time Lag

Early Laboratory Research

In 1957, Joe Adams published a paper using an eleven-point "confidence scale" with a zero-alternative task. His subjects were trained to use, not probabilities, but a scale defined to them in precisely the way we have defined calibration: "[the subject was] instructed to express his confidence in terms of the percentage of responses, made at that particular level of confidence, that he expects to be correct. . . Of those responses made with confidence p, about p% should be correct" (p. 432-3).

Each of forty words was presented tachistoscopically ten times successively, with increasing illumination each time, to ten subjects. After each exposure subjects wrote down the word they thought they saw, and gave a confidence judgment, limited to the numbers 0, 10, 20, . . . 90, 100. The resulting calibration curve, across subjects, is shown in Figure 4. Great caution must be taken in interpreting the data: because each word was shown 10 times, the responses are highly interdependent. It is unknown what effect such interdependence has on calibration, but the finding of gross underconfidence along the entire response scale has been replicated with only one subject in one experiment (Swets, Tanner and Birdsall, 1961). Perhaps subjects were "holding back," unwilling to give a high response when they knew that the same word would be presented several more times.

The following year Adams and Adams (1958) reported a training experiment, using the same response scale, but a new, three-alternative, single-response task: For each of 156 pairs of words per session, subjects were asked whether the words were antonyms, synonyms, or unrelated. Thirteen of the 14 experimental subjects, who were shown calibration tallies and calibration curves after each of five sessions, had lower discrepancy scores on the fifth day than on the first. The mean decrease for the 14 subjects was 48%. Six control subjects, whose only feedback was a tally of their unscored responses,

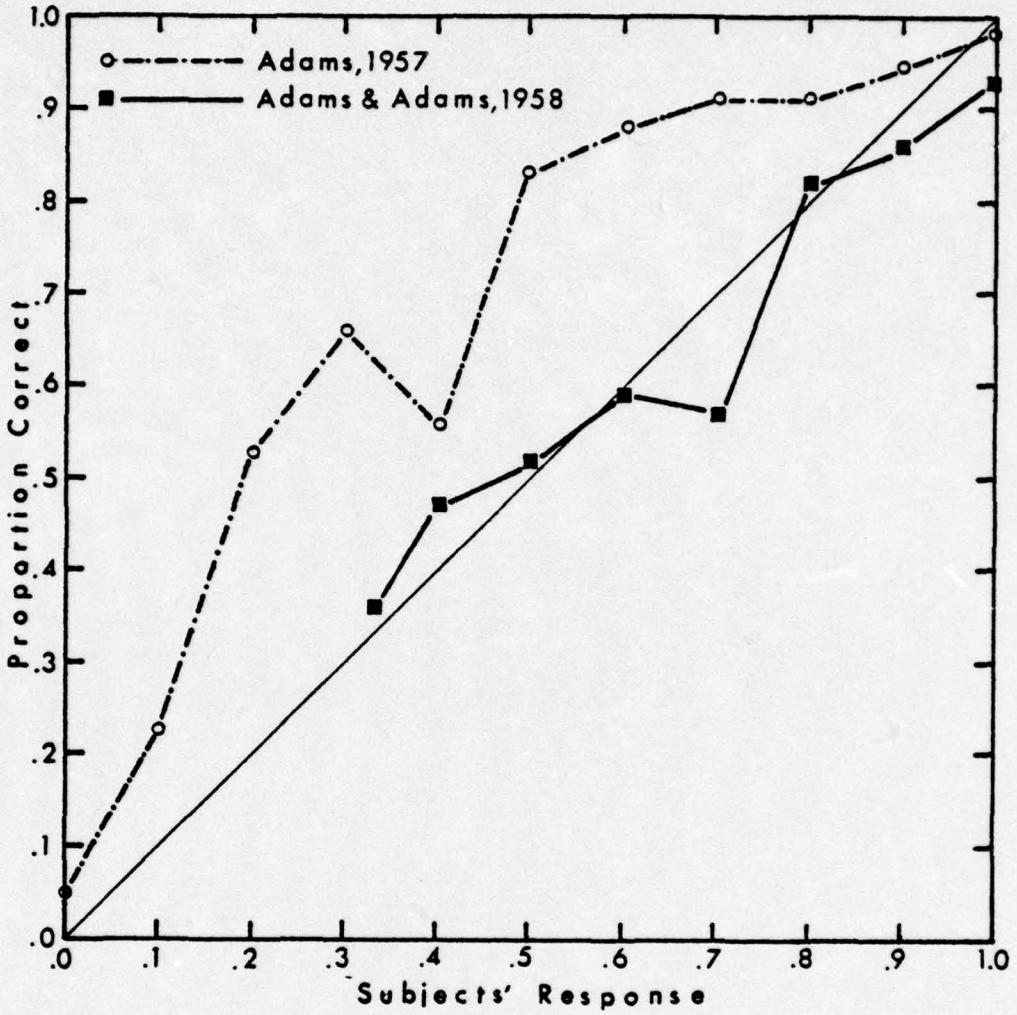


Figure 4

Calibration from the Adamses' Data

showed a 36% mean increase in discrepancy scores. Figure 4 shows the calibration, grouped across all five sessions for one experimental subject--the only subject for whom Adams and Adams reported such data.

In a 1961 Psychological Review article, Adams and Adams discussed many aspects of the calibration of probabilities (using the term "realism of confidence"), anticipating much of the work done by others in recent years, and presented more bits of data, including the grossly overconfident calibration curve of a schizophrenic who believed he was Jesus Christ. They reported calibration curves from a nonsense syllable learning task with large overconfidence after one trial and improvement after 16 trials. They also described briefly a "transfer of training" experiment: On the first day, subjects made 108 decisions about the percentage of blue dots in an array of blue and red dots. On the second and fourth days, the subjects decided on the truth or falsity of 250 general statements. On the third day, they lifted weights blindfolded. On the fifth day, they made 256 decisions (synonym, antonym, or unrelated) about pairs of words. Eight experimental subjects, given calibration feedback during the first four days, showed on the fifth day a mean absolute discrepancy score significantly lower than that of eight control (no feedback) subjects, suggesting some transfer of training. Finally, Adams and Adams reported a correlation of .36 between absolute discrepancy scores and fear of failure (achievement anxiety) for 56 subjects taking a multiple choice final examination in elementary psychology. Neither over- nor underconfidence nor knowledge was related to fear of failure, only calibration.

One can suppose that, having originated such a wide range of thoughtful ideas, the Adamses sat back to watch the procession of further work on the topic. If so, they may still be waiting. Except for the study by Oskamp (1962) described next, no other work appeared for over ten years, and of all

the other literature reviewed in this paper, not a single author has referenced the Adamses' or Oskamp's work!

Oskamp (1962) used 200 MMPI profiles[3] as stimuli. Half the profiles were from men admitted to a VA hospital for psychiatric reasons; the others were from men admitted for purely medical reasons. The task was to decide, for each profile, whether the patient's status was psychiatric or medical, and state the probability that the decision was correct, using the half-range method. Each profile had been independently categorized as hard (61 profiles), medium (88), or easy (51) on the basis of an actuarially-derived classification system, which correctly identified 57%, 69%, and 92% of the hard, medium, and easy profiles.

Three groups of subjects judged all 200 profiles: 28 undergraduate psychology majors, 23 clinical psychology trainees working at a VA hospital, and 21 experienced clinical psychologists. The 28 inexperienced judges were later split into two matched groups, and given the same 200 profiles again. Half were trained to improve accuracy: after the first 50 repeated profiles, they were told their percent correct for the first 200 and the just-completed 50, and instructed in the use of four simple actuarial rules (e.g., if the F-scale is 55 or higher, call the profile psychiatric). For profiles 51 through 100, they received right/wrong feedback after every 10 profiles. They received no feedback during profiles 101-200. The other inexperienced judges received calibration training during their second session. After every 50 profiles, they were told their percent correct, their calibration score, their rank within the group on both these measures, and shown their calibration curve. The experimenter suggested and discussed

---

[3]The MMPI (Minnesota Multiphasic Personality Inventory) is a personality inventory widely used for psychiatric diagnosis. A profile is a graph of 13 sub-scores from the inventory.

ways of improving each subject's calibration.

Oskamp used three measures of subjects' performance: accuracy (percent correct), confidence (mean probability response), and appropriateness of confidence (a calibration score: $\frac{1}{N}\sum_t n_t|r_t-\bar{c}_t|$). All three groups were, in general, overconfident, especially the undergraduates in their first session (accuracy 70%, confidence .78). However, all three groups were underconfident on the easy profiles (accuracy 87%, confidence .83).

The subjects trained for accuracy increased their accuracy from 67% to 73%, closer to their confidence, .78, which did not change as a result of training. Their calibration score decreased from .17 to .10.[4] The subjects trained for calibration lowered their confidence from .78 to .74, bringing it closer to their accuracy, .68, which remained unchanged. Their calibration score decreased from .15 to .11.

Signal Detection Research

In the early days of signal detection research, investigators looked into the possibility of using confidence ratings rather than Yes-No responses in order to reduce the amounts of data required to determine a stable ROC (receiver operating characteristic) curve. The classic Psychological Review paper by Swets, Tanner and Birdsall (1961, the same volume in which the Adamses' review appeared!) reported individual calibration curves for four observers who used a six-point rating scale to indicate their confidence that they had heard a signal plus noise rather than noise alone. The ratings were defined on a probability scale, the first point representing 0.0 to 0.04, the next 0.05 to 0.19, followed by four equal-width categories, 0.20–0.39, 0.40–0.59, 0.60–0.79, 0.80–1.00. The calibration curves of the four subjects,

[4] MMPI-buffs might note that with this minimal training the undergraduates showed as high an accuracy as either the best experts or the best actuarial prediction systems.

based on 1,200 trials each, are shown in Figure 5. The individual differences are striking, with only one subject being even remotely well calibrated.

Clarke (1960) reported an experiment in which one of five different words, mixed with noise, was presented to listeners through headphones. The listeners selected the word they thought they heard and then rated their confidence by indicating one of five categories defined by slicing the probability scale into five ranges. Twelve practice tests of 75 items each helped the listeners to calibrate themselves. After each test, listeners scored their own results and noted whether the appropriate percentage of correct identifications fell in each rating category, thus allowing them to change strategies on the next test. Clarke found that although all five listeners appeared well calibrated when data were averaged over the five stimulus words, analyses for individual words showed that the listeners tended to be overconfident for low-intelligibility words and underconfident for words of relatively high intelligibility. As we show in the next section, this pattern of findings, overconfidence for difficult items and underconfidence for easy items, has been obtained in different tasks.

Clarke also reported an experiment in which both the signal-to-noise ratio and the number of alternatives were varied. He found that the calibration curves for different signal-to-noise ratios were nearly identical when only four words made up the message set. But when any one of 16 words was possible, the curves appeared well calibrated only for the larger signal-to-noise ratios, deteriorating of overconfidence at smaller signal-to-noise ratios. In spite of their training in using the rating scale, the listeners adopted different response criteria for different stimulus characteristics, thereby shifting their calibration curves.

Pollack and Decker (1958) used a verbally defined 6-point confidence rating scale that ranged from "Positive I received the message correctly" to "Positive I received the message incorrectly." With this rating scale

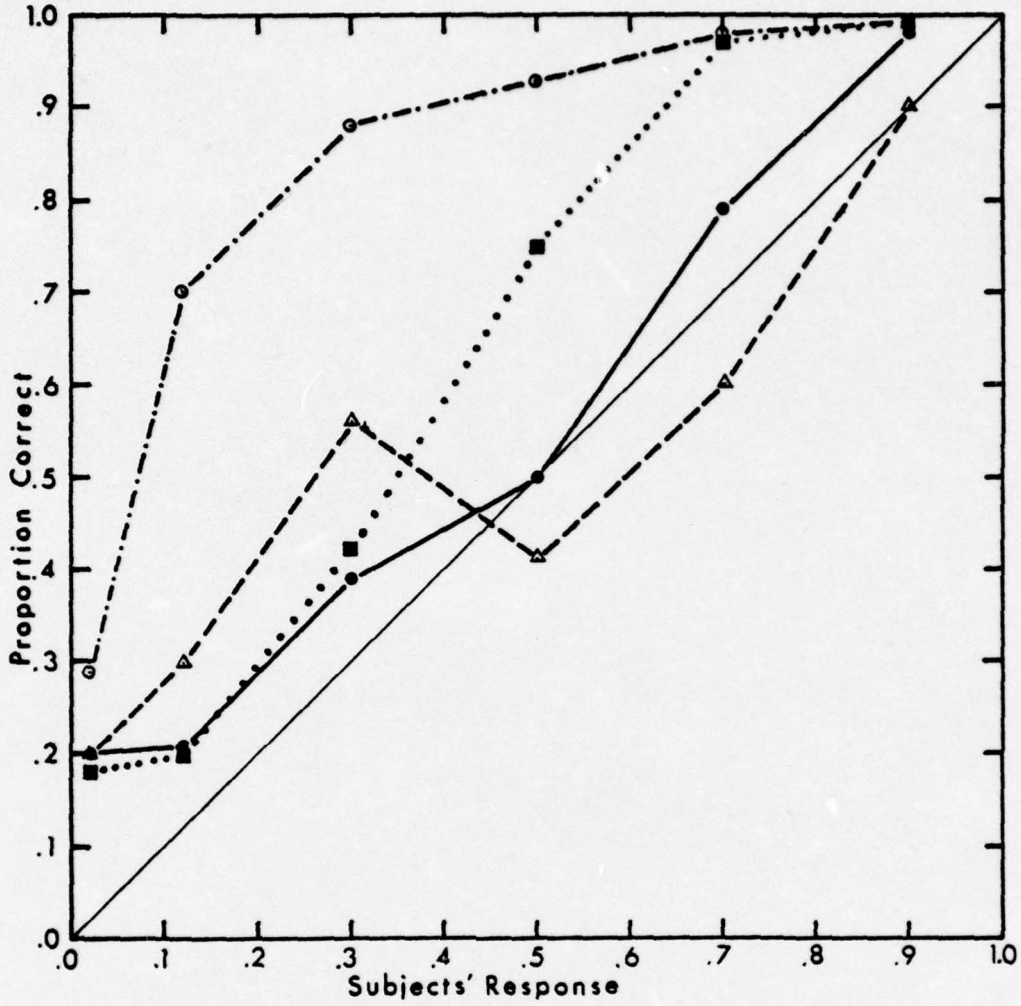


Figure 5

Calibration for Four Subjects in a Signal Detection Task

it is impossible to determine whether an individual is well calibrated, but it is possible to see shifts in calibration across conditions. In seeming contrast to Clarke's results, Pollack and Decker showed that the average calibration curve over three subjects remained unchanged with different signal-to-noise ratios. However, when subsets of difficult items, medium items and easy items were analyzed separately, the invariance of the calibration curves disappeared. Calibration curves for easy words generally lay above those for difficult words, whatever the signal-to-noise ratio, and the curves for high signal-to-noise ratios lay above those for low signal-to-noise ratios, whatever the word difficulty.

In another experiment on message reception, Decker and Pollack (1958) varied the frequency cutoffs for the noise that was mixed in with the speech. For one subject, calibration was unaffected by the change in filters, but for the other two subjects, the calibration curve for the lower-frequency filter was below that for the other filter. Here, the effect of task difficulty on calibration depended on the individual.

In most of these studies, shifts in calibration curves were of secondary interest; the important question was whether confidence ratings would yield the same ROC curves as Yes-No procedures. To answer this question, it is not necessary to define rating scales in terms of probabilities; verbally-defined categories are sufficient. Thus, the probability scale disappeared from signal detection research. By 1966, Green and Swets concluded that, in general, rating scales and Yes-No procedures yield almost identical ROC curves. Since then, studies of calibration have disappeared from the signal detection literature.

Recent Laboratory Research

Hazard and Peterson (1973) found no effect on calibration due to

changes in response mode. Forty subjects, armed forces personnel studying at the Defense Intelligence School, responded with probabilities, and with odds, to 50 two-alternative general knowledge items (e.g., which magazine had the largest circulation in 1970, Playboy or Time?), using the half-range method. Substantial overconfidence was found, as shown in Figure 6. Lichtenstein (unpublished) replicated the results, using the same items but only the probability response, with 19 Oregon Research Institute employees. Phillips and Wright (in press) found similar results with different items, using British undergraduate students as subjects. The calibration curves shown in Figure 6 look remarkably similar considering the variety of subject populations employed; all showed gross overconfidence.

Using the same half-range, two-alternative method, we have recently conducted a series of experiments exploring calibration (Lichtenstein and Fischhoff, 1976). We will briefly review our findings here.

In two tasks chosen to be extremely difficult, subjects were poorly calibrated; in fact, they showed no evidence of calibration at all. Figure 7 shows curves for these tasks, one in which subjects were asked to identify small sketches as drawn by European or Asian children, and one in which they studied stock market charts and were asked to predict whether the stock described by each chart would be up or down 3 weeks hence. Overall percent correct was 53% for children's art, 47% for stocks.[5]

Even a small amount of substantive knowledge will induce some improvement in calibration. We asked two other groups of subjects whether each of 10 examples of handwriting was written by a European or an American, after they had studied 10 similar examples. All examples were preselected to be difficult to judge. The training group's study examples were correctly

[5] We caution the reader against trying to interpret the fascinating shape (a fish?) created by these two calibration curves. We think it's a fluke of chance.

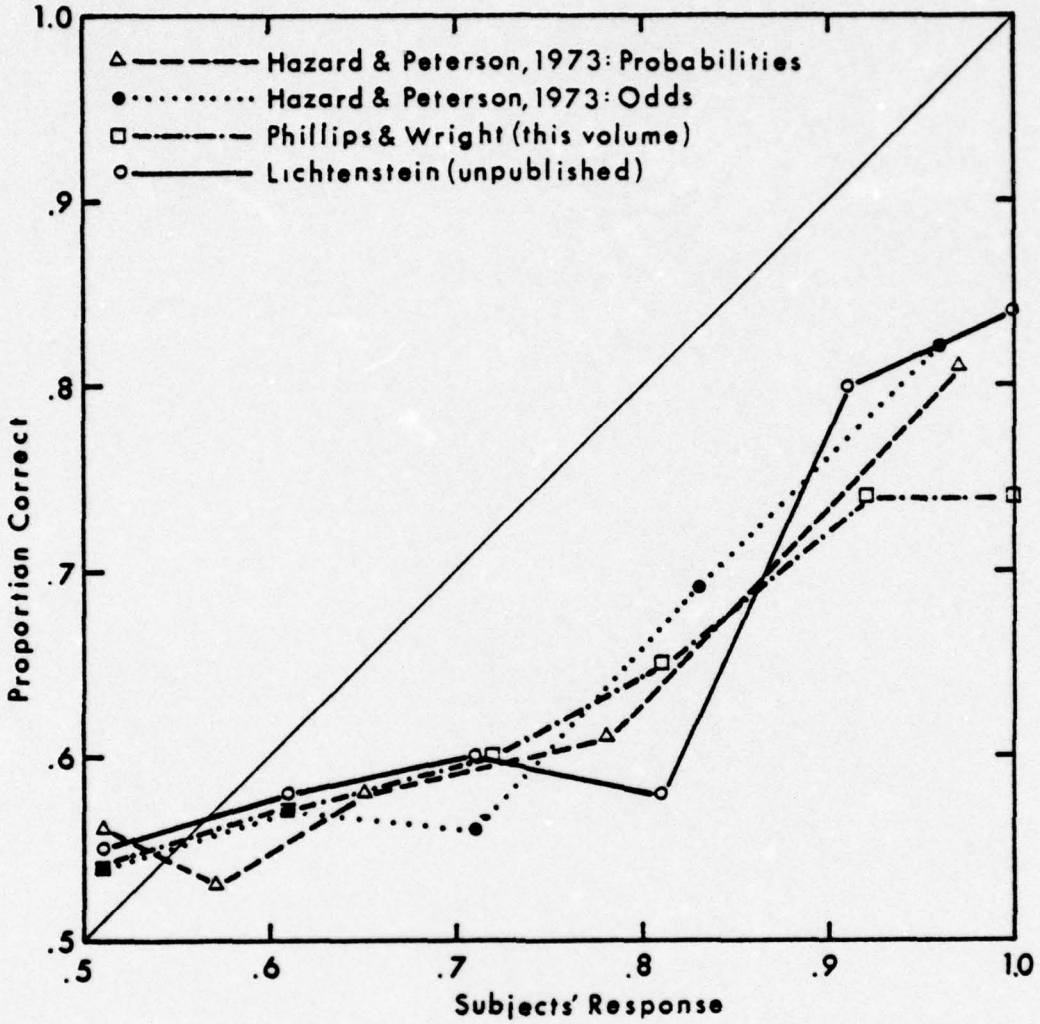


Figure 6

Calibration for Half-Range Tasks

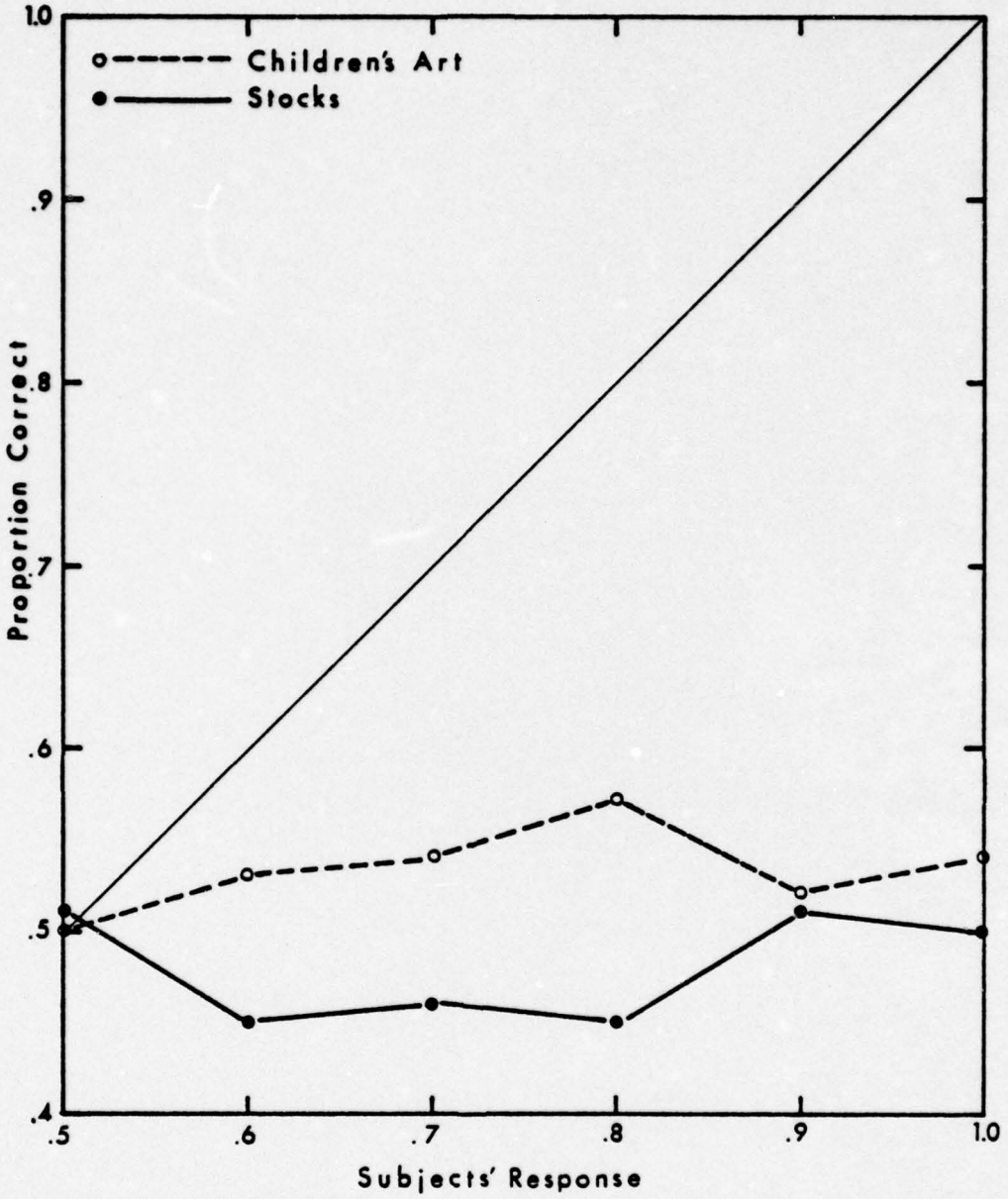


Figure 7

Calibration for Two Impossible Tasks

labeled as to country of origin; the no-training group's study examples were unlabeled. As shown in Figure 8, the training group, who correctly identified 71% of the handwriting examples, were much better calibrated than the no-training group (51% correct).

We pursued the notion that substantive knowledge affects calibration in several additional studies using two-alternative general knowledge items. Substantive knowledge was defined for subjects by the proportion of items they correctly answered (best or worst subjects) and for items by the proportion of correct answers, across subjects, for each item (easy or hard items). Figure 9 gives results for 50 graduate students pursuing Ph.D.'s in psychology. A replication using different items and a different sample of subjects, undergraduate student volunteers, showed similar results (not graphed here).

These curves clearly show that the degree of over- or under-confidence is a function of substantive knowledge. The most knowledgeable subjects *answering the easiest items* showed substantial underconfidence, while the worst subjects on the hardest items showed substantial overconfidence. The relationship between item difficulty and over- or under-confidence is mediated by the distribution of responses given by subjects. To be well calibrated with hard items, an assessor must use many responses of .5 and .6 and a few of .9 and 1.0, while with easy items the reverse must be true to achieve good calibration. The distributions of responses for the four calibration curves shown in Figure 9 indicate that the subjects did change their distributions, but not as much as they should have. Across 16 different experiments or sub-experiments we have run (Lichtenstein and Fischhoff, 1976) using two-alternative half-range tasks, there is a .91 correlation between the mean response over all subjects and items (range .65 to .86) and the percent correct over all subjects and items (range 43 to 92), giving further

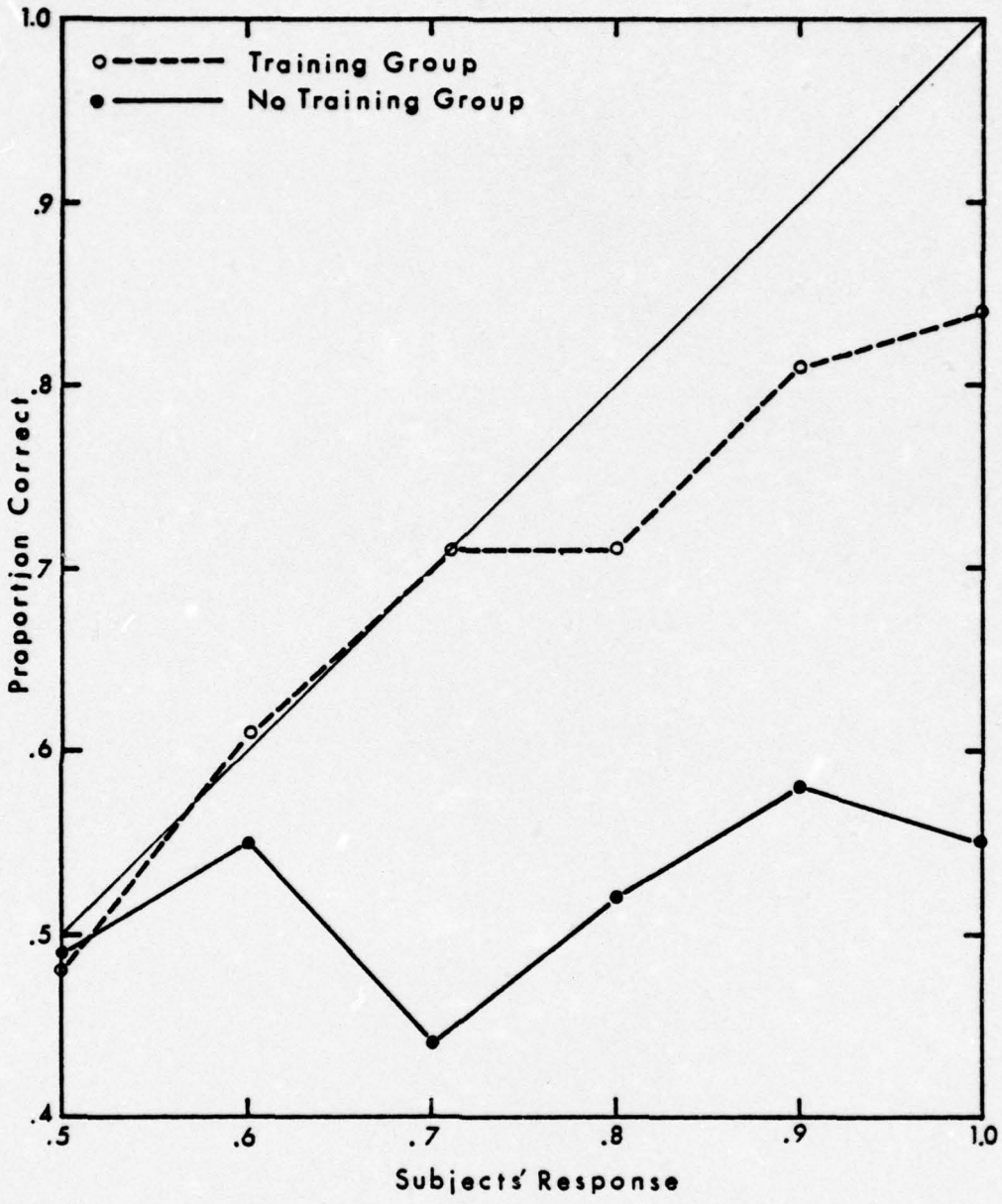


Figure 8

Calibration for Handwriting
Identification: Training versus No Training

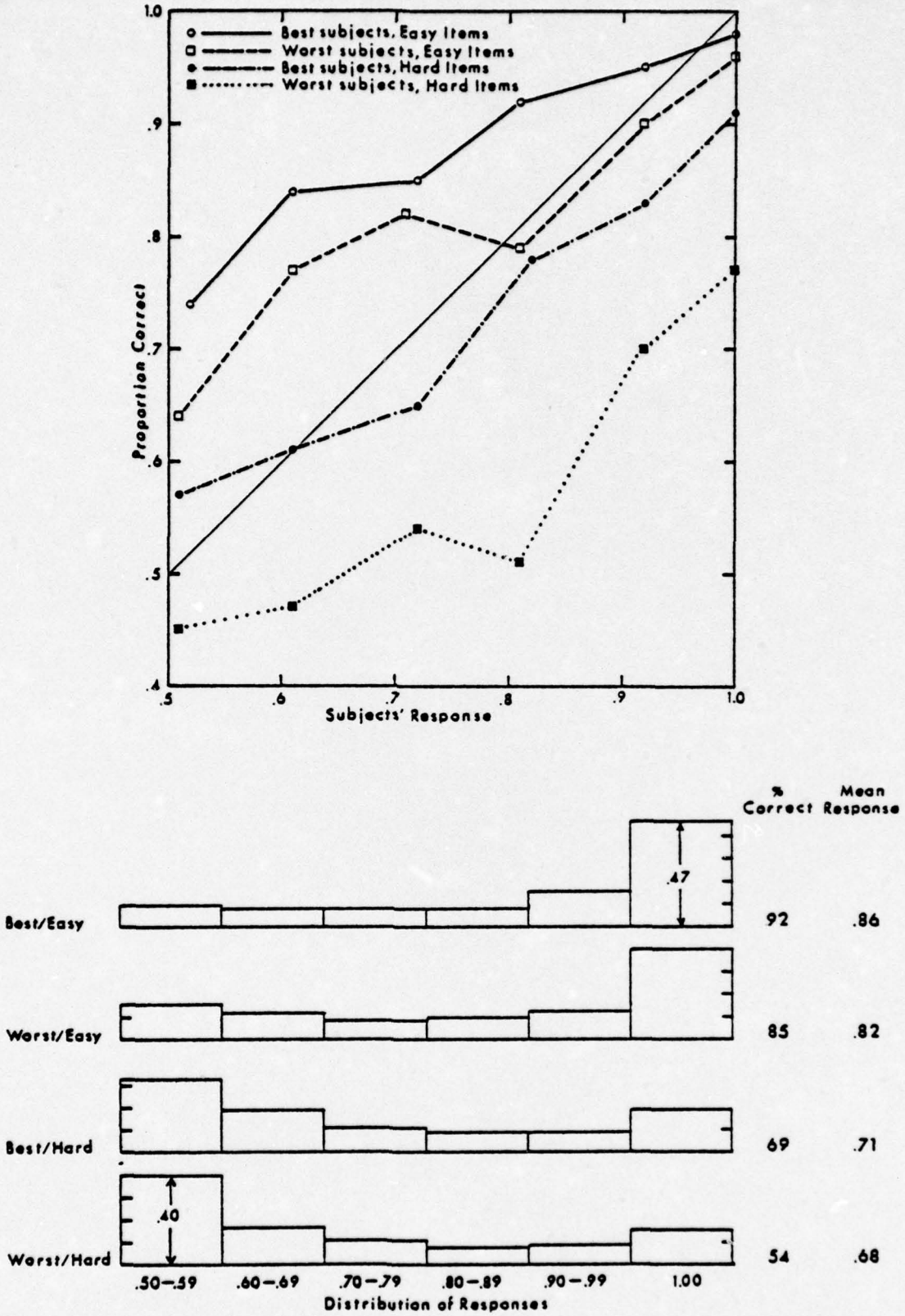


Figure 9

Calibration for Subsets Varying in Difficulty

evidence that subjects do change their response distributions as the difficulty level of the task changes, though not enough to achieve good calibration.

The calibration curves shown in Figure 9 were not calculated from separate, independent sets of data, but from subsets of items embedded in a larger set, the longer test given to each subject. To guard against the possibility that there is some artifactual reason for these findings, due perhaps to an adaptation level effect operating in the larger, more varied tests the subjects actually took, we prepared two tests, one hard (50 items) and one easy (50 items), using items that had previously been used in a large, varied test. These smaller tests were given to two new groups of subjects; 48 subjects took the hard test, 45 the easy. Figure 10 shows that the calibration from these two separate, independent tests was essentially the same as calibration calculated from sub-tests created artificially (and post hoc) from a larger set of data. The effect of test difficulty shown here is not an artifact due to our method of analysis.

Using a full-range, one-alternative task, Pitz (1974) found an item-difficulty effect similar to that reported above. He gave 38 subjects 12 items concerning the population of various countries (e.g., "the population of Brazil exceeds 85 million"), and an unspecified number of items concerning the grade each would receive in Pitz's course, one week before the final exam. The population items were chosen to be difficult, the course grade items easy. The divergence of the two calibration curves is apparent (see Figure 11).

While Pitz did not report percent correct for either group, his "hard item" calibration curve is similar to data Fischhoff and Lichtenstein (in preparation) have collected with the two-alternative, full-range method (see

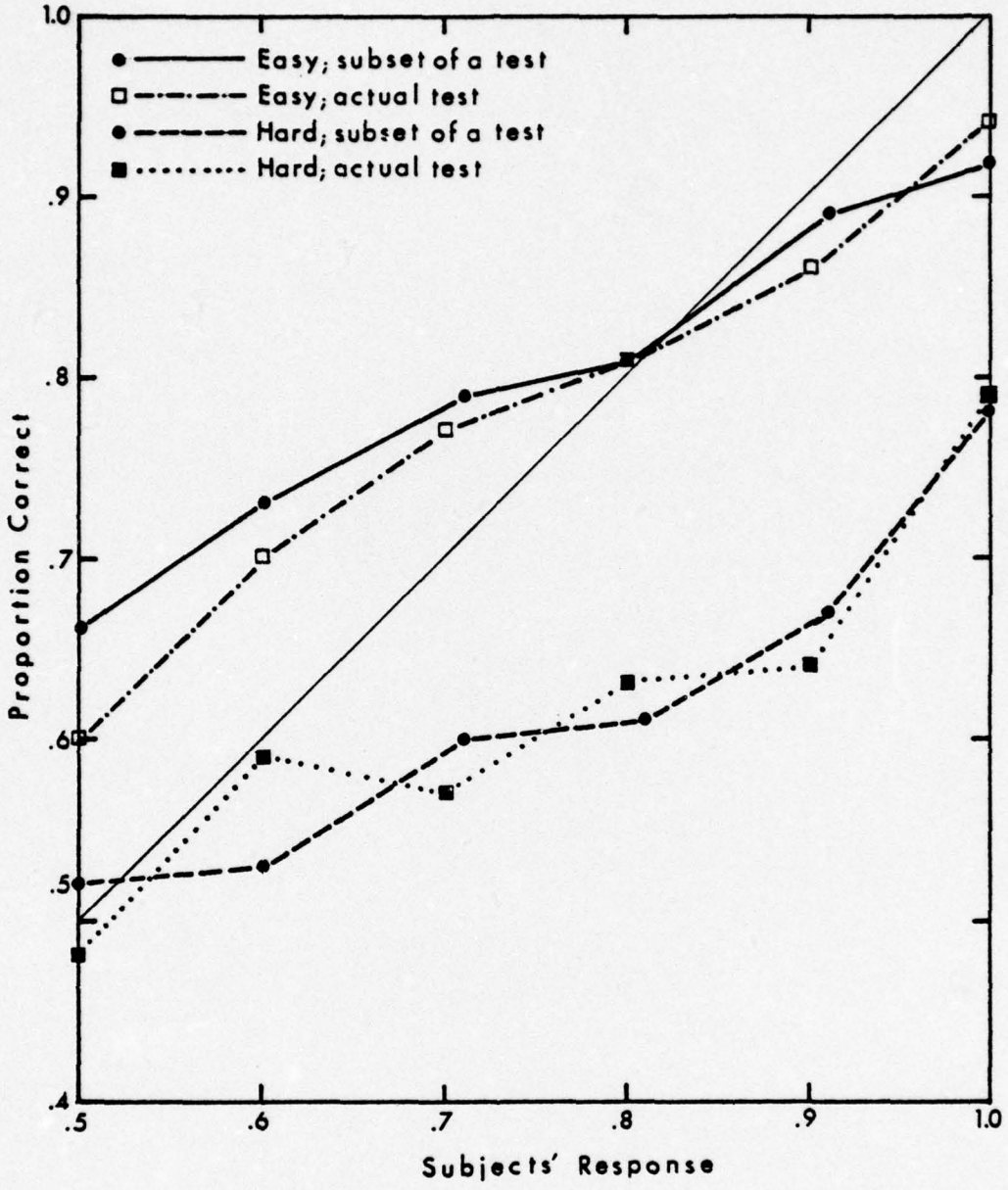


Figure 10

Calibration for Hard and Easy Tests Versus
Hard and Easy Subsets of a Test

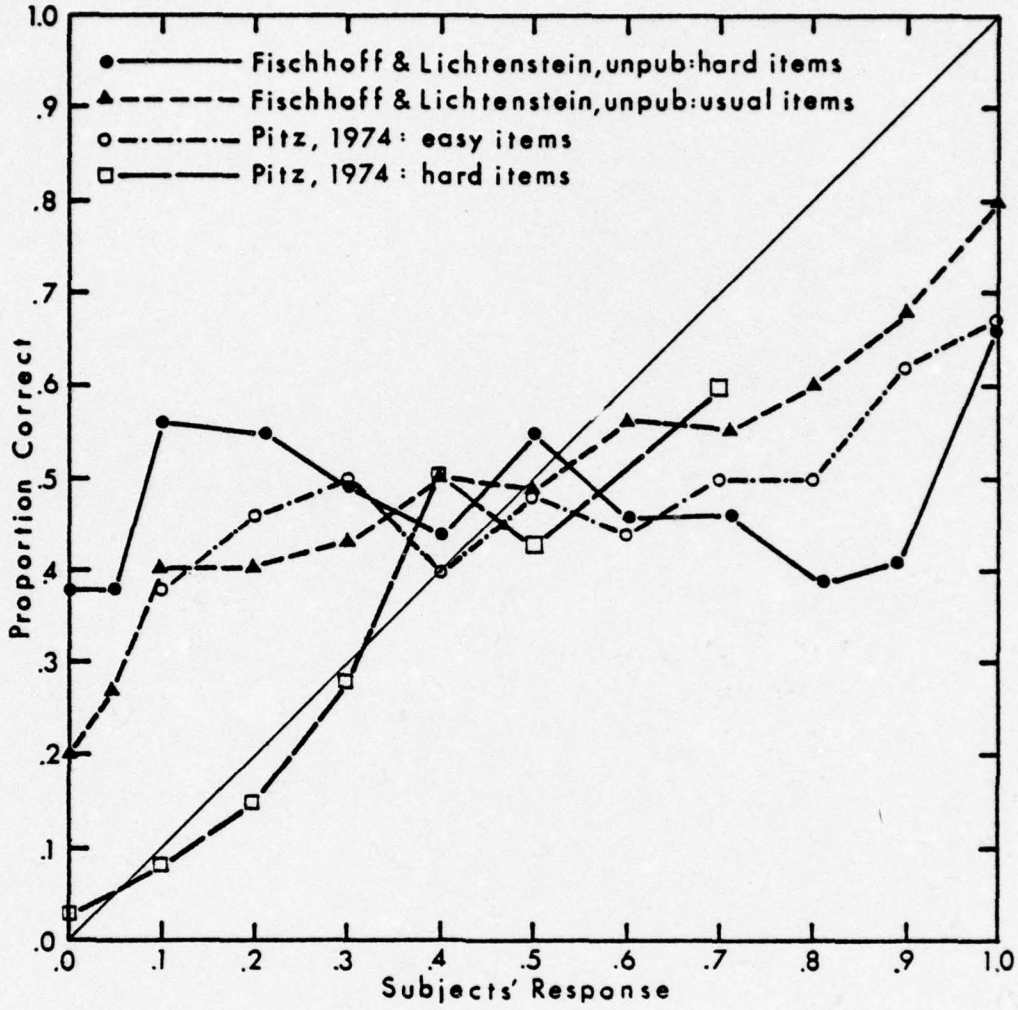


Figure 11

Calibration for Several Full-Range Studies

Figure 11). In our study, 100 two-alternative items were given to 131 subjects. Half the subjects were told to assess the probability that the first alternative was correct; the other half responded to the second alternative. The data from the two groups were combined. The test items were composed of two subsets, one with 75 items of moderate difficulty (65% correct)[6] and one with 25 items of greater difficulty (55% correct). Clearly, the pattern of Pitz's results for hard items was repeated; the calibration was abysmal.

Perhaps the categorization of items into "hard" and "easy" does not really capture the essence of expertise. Experts might be better calibrated not only because they know the correct answer for more of the items, but also because they have thought more about the whole topic area, and thus can more readily recognize the extent and the limitations of their knowledge. We tested this hypothesis, using psychology graduate students as our experts. They responded to 100 items, 50 dealing with knowledge of psychology and 50 dealing with general knowledge. The two parts of the test were analyzed separately. The percent correct was the same (76%) for the two parts. Since item difficulty was controlled for, differences in calibration could only be attributed to the hypothesized quality of insight that experts might have above and beyond their level of knowledge. As shown in Figure 12, no such differences were found.

---

[6] In the full-range method, percent correct is calculated as follows: when the subject responds with a probability > .5, we count the successes; when the response is .5, we count half the responses, under the assumption that the subject, when asked to choose which of two alternatives is the preferred one, would randomly make that choice. When the response is < .5, we count the failures: if you say the probability of rain tomorrow is .1, and it doesn't rain, then you were correct in believing it would more likely <u>not</u> rain than rain.

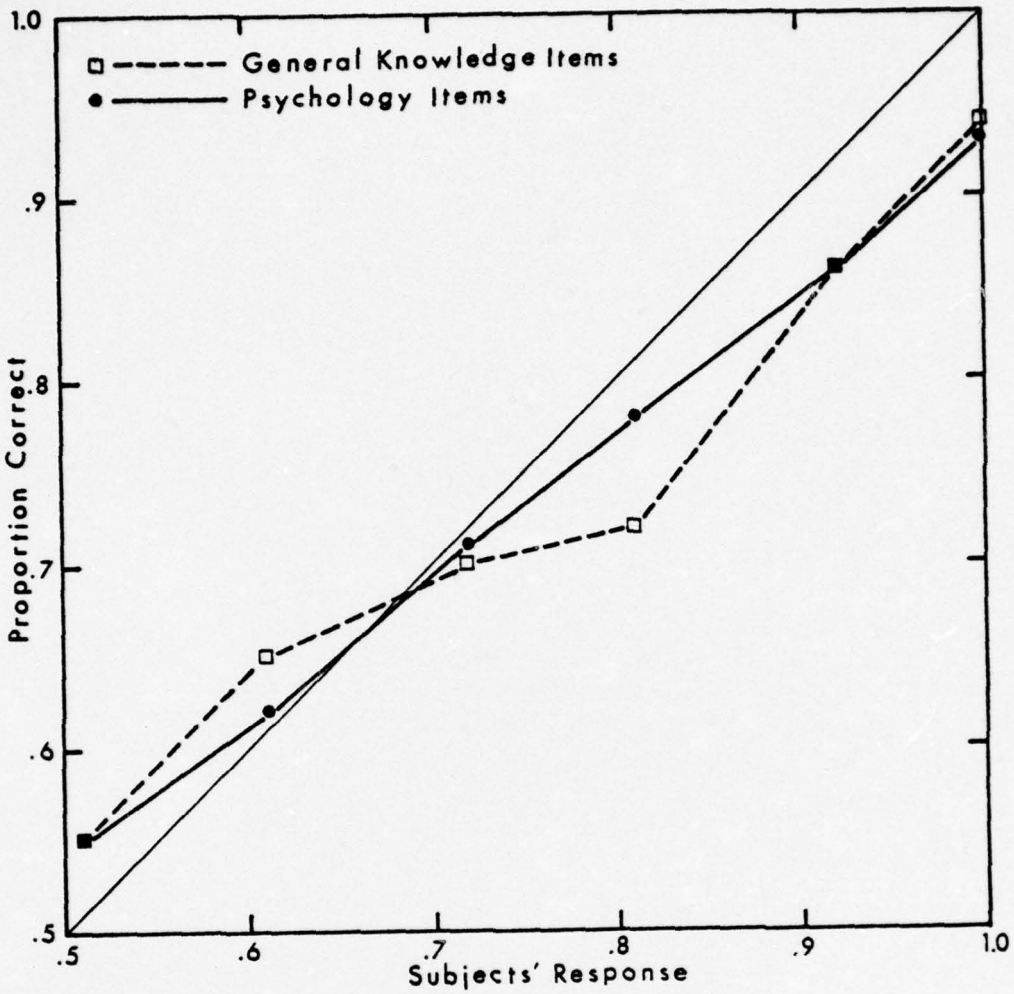


Figure 12

The Effects of Special Topical Knowledge

Finally, we looked at the effect of intelligence on calibration. Our usual volunteers were mostly undergraduate college students. Our graduate student subjects may be presumed to be significantly more intelligent, as a result of highly selective admissions procedures. Figure 13 shows the calibration of two subtests of 73 items. The subtests were chosen from previously collected data so that each item from the usual volunteers was matched in difficulty (% correct) by an item from the graduate students. The graduate students appear to be slightly better calibrated at .5 and 1.0. The differences are slight, however, when compared with differences in calibration due to test difficulty.

Data from two full-range studies are shown in Figure 14. Fischhoff and Beyth (1975) asked 150 Israeli university students to assess the probability of 15 then-future events, possible outcomes of President Nixon's much-publicized trips to China and Russia. Examples of the events are "President Nixon will meet Mao at least once"; "The USA and the USSR will agree to a joint space program"; "President Nixon will announce that his trip was successful." The resulting calibration curve, based on 1,921 assessments, is suboptimal at 0 and 1, and shows a dip at .7, but is otherwise remarkably close to the identity line. Why? The subjects received the usual instructions. They were not experienced in probability assessments. They were run in large classroom groups. They were not foreign-affairs experts. Is this ability a special attribute of Israelis?

Sieber (1974) had 20 subjects assess probabilities for all four multiple-choice alternatives of 20 items in a college classroom exam. All 1600 responses are included in this curve. A large proportion of the responses (77%) were of the form (1, 0, 0, 0), and for these responses the calibration was superb: the percent correct was 98.7. The rest of the curve (see Figure 14) is based on few data. It is difficult to know to

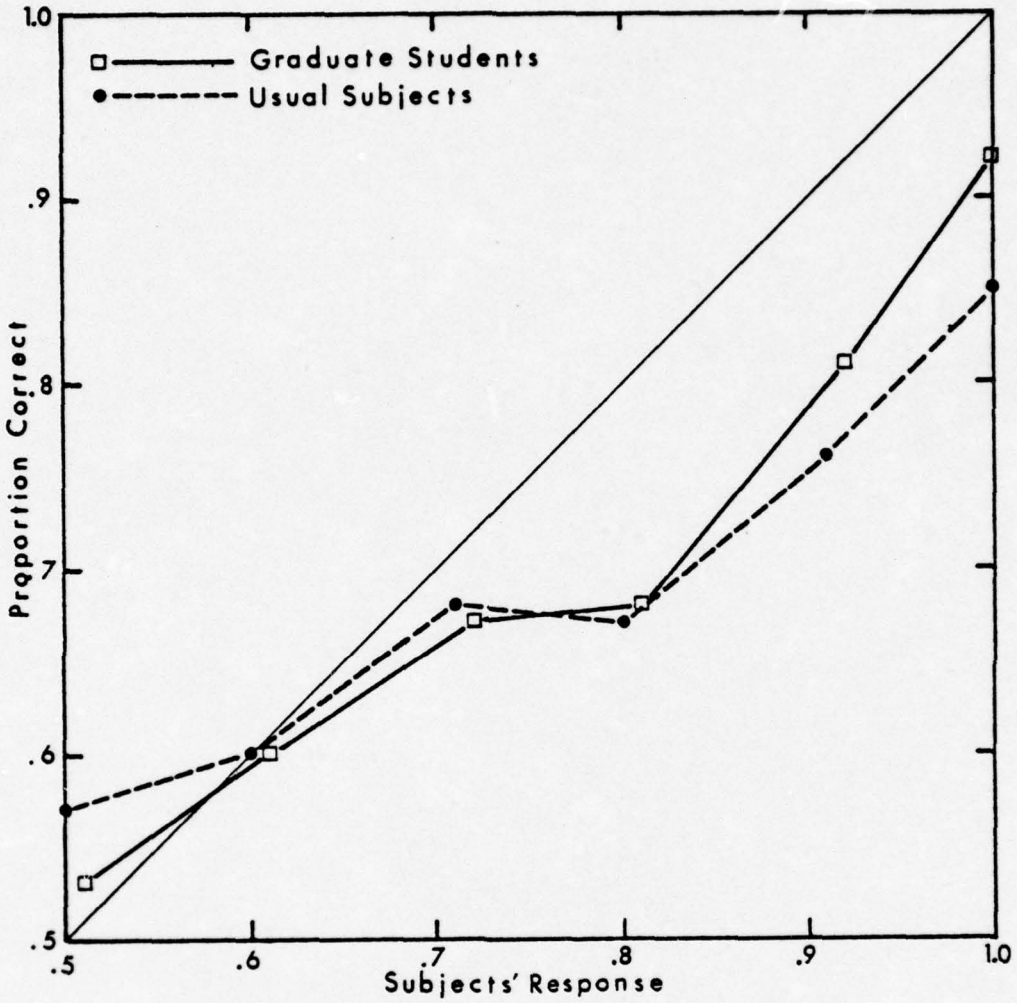


Figure 13

The Effects of Intelligence

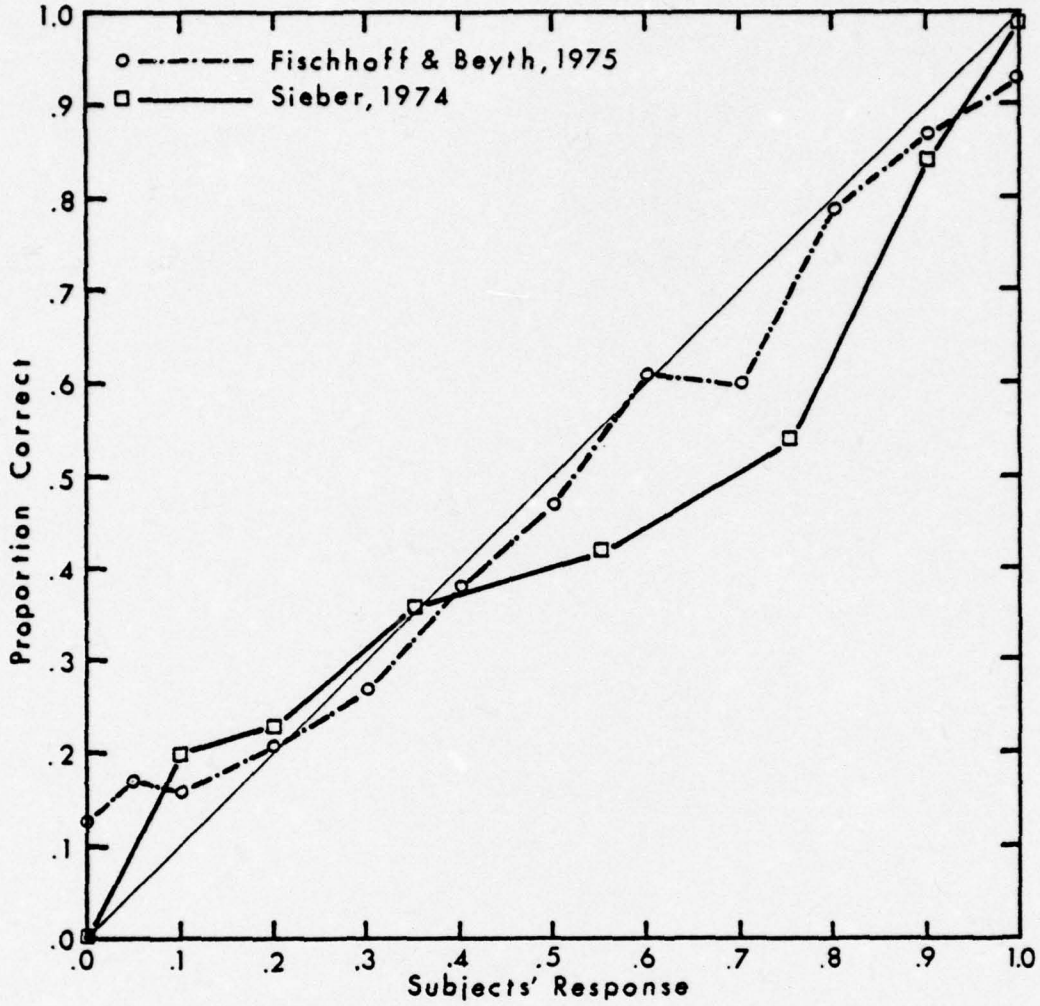


Figure 14

Two Full-Range Studies

what extent the apparent symmetry about the point (1/4, 1/4) is forced on the curve by the inclusion of all four responses to each item.

The primary purpose of Sieber's experiment was to study the effect of motivation on calibration. The subjects whose data are plotted here were told that the score they earned on the test (based on a proper scoring rule) would not count in their grade. Another group was told their score would count in their grade. The latter (highly motivated) group used (1, 0, 0, 0) for 90% of their responses. Their calibration (not plotted here) appears worse, but so little data are available for the curve (aside from the end points) that one should be cautious in drawing any conclusion.

In a stock market prediction task, Staël von Holstein (1972) asked subjects to assess probabilities for a five-alternative task: the future movement of stocks categorized into five intervals fixed by the experimenter. He did not report the data necessary to compute a calibration curve, except to note, tantalizingly, that of 7,896 distributions only 40 were of the extreme form (1, 0, 0, 0, 0). Of these, only 12 were correct!

The full-range studies based on laboratory research, shown in Figures 11 and 14, indicate symmetric calibration: the proportion correct for any response r is approximately equal to one minus the proportion correct for the response 1-r. In contrast, the full-range calibration curves from the weather forecasting studies shown in Figures 1 and 2, are not (except for Root, 1962) symmetric: they show a constant bias across the entire range. It is tempting to believe that whether a calibration curve shows symmetry or bias depends on the implicit payoff structure for different kinds of error. Forecasters may prefer to forecast rain and be wrong than to forecast no rain and be wrong. But it seems unlikely that laboratory subjects perceive differential penalties for saying absinthe is a liqueur and finding out it is a precious stone versus saying it is a precious stone and finding out it is a liqueur.

A rarely-discussed problem in measuring an assessor's calibration is the large number of assessments needed to provide a stable estimate. One way to reduce the number of responses required is to assume that the calibration curve is one of a family of curves, and use the data to estimate the parameters of the curve. Shuford and Brown (1975; see also Brown and Shuford, 1973) assumed that calibration curves are straight lines, and found least squares estimates of the slope and intercept for each subject. The model becomes a one-parameter (slope) model when, for n items with k alternatives, the subject gives responses to all alternatives and all nk responses are fitted by the model. Provided that the sum of the k responses to a single item is always 1.0, the fitted line is constrained in their model to pass through the point (1/k, 1/k). Using 3-alternative items, Shuford and Brown reported, without supporting detail, that "as long as a reasonably wide range of [responses] is used by the [subjects], this estimation procedure can yield fairly stable results with 15- and 20-item tests" (1975, p. 157). However, the authors were concerned that their model assumes that all responses are independent, and suggested that when more than two alternatives are used, this might not be true because "some people might tend to overvalue information when deducing reasons in favor of an answer, but tend to undervalue information when deducing reasons against an answer" (p. 157). To solve this problem, they proposed a planar least-squares estimation procedure for the special case of three alternatives. The planar model, however, did not produce stable estimates for small numbers of items.[7]

Schlaifer (1971), in his MANECON program called TRUCHANCE, proposed a one-parameter model which is linear in the log of the odds of the response (r) plotted against the log of the odds of the proportion correct (c):

$$\log \frac{c}{1-c} = A + \log \frac{r}{1-r}$$

---

[7] T. A. Brown, personal communication, March 3, 1975.

His program uses a Bayesian approach to finding the posterior distribution of the parameter A, given a set of responses, and uses that distribution to recalibrate future responses. This model is somewhat limited. The only forms of miscalibration it can recognize are curves always above the diagonal or always below it. Such a model could not adequately represent the symmetric full-range data shown in Figure 1 (Root, 1962) and Figure 11.

We have recently been exploring the use of models to improve the stability of estimates of calibration (Phillips and Lichtenstein, in preparation), using both a two-parameter linear model and a two-parameter expansion of Schlaifer's model:

$$\log \frac{c}{1-c} = A + B \log \frac{r}{1-r} \quad .$$

We are less sanguine than Shuford and Brown about the number of items required for stable estimation. Consider an assessor who is so badly calibrated that she says .2 when she ought to say .35, and says .8 when she ought to say .7. Preliminary results with simulated data indicate that the probability that such an assessor will <u>appear</u> to be perfectly calibrated can be as high as .5 for a 100-item test.

The need for accurate estimates of calibration with the fewest possible data is most pressing when one considers the problem of training an assessor to become better calibrated. An obvious design for a training experiment would be to run a subject for, say, eight sessions. At the end of each session, we would give her feedback, telling her about her calibration and urging her to improve it. If we collect too few data per session, we stand a large chance of giving her false feedback--telling her, for example, that she is consistently underconfident, when in fact she is really overconfident. In addition, the experimenter in such a study would have little power (in the statistical sense) to conclude, after the experiemnt, that training led

to improvement. On the other hand, preparing and presenting 800 to 1600 stimuli (100 to 200 per session) presents problems for both the experimenter and the subject.

Brown and Shuford (1973) have suggested two ways of dealing with this problem: (1) Give subjects scoring-rule feedback after every item. This might serve to keep subjects interested and learning. (2) Give calibration feedback after every N items. This feedback would be the straight line fitted to the data. They further suggest that all responses to each item, not just one response, be fitted. We believe that using all the data might work for those situations where a constant bias is unlikely, such as when using diversified items of general information. But when the items are repeated presentations of the same question, such as "Will it rain tomorrow?", the inclusion of both responses to each item would tend to obscure the kind of bias shown in Figures 1 and 2.

One further problem in training assessors is the possibility that the assessor will trade off information transmission for calibration. At the extreme, an assessor could always respond with the base rate (the overall proportion of correct propositions), thus yielding excellent calibration but no information. To avoid this strategy, it might be wise to feed back to the trainee Murphy's vector partitions of the scoring rule (or, where appropriate, the special scalar partitions) at the end of every session. Hopefully, the subject would learn to improve the calibration portion of the score without greatly decreasing the resolution portion. In addition, one would wish to show the trainee, perhaps via a calibration curve smoothed by a fitted model, whether poor calibration was due to overconfidence or underconfidence.

Our previous finding that subjects tend to be overconfident with hard items and underconfident with easy items adds to the dilemma one faces in

planning a training experiment. Those data suggest that one might have to train subjects in both hard and easy tasks, separately, to have any hope that the training would generalize.

CONTINUOUS PROPOSITIONS: UNCERTAIN QUANTITIES

Continuous uncertain quantities can be proportions (What proportion of students prefer Scotch to Bourbon?) or numbers (What is the shortest distance from England to Australia?). Subjects are usually not asked to draw the entire density function across the range of possible values. The elicitation procedure most commonly used is some variation of the fractile method. In this method, the subject is asked to give the median of the distribution ("state a value such that the true value is equally likely to fall above or below the value you state"), and then several other fractiles. For example, for the .01 fractile the subject would be asked to state a value such that there is only 1 chance in 100 that the true value is smaller than the stated value. In one variant called the tertile method, the subject is not asked the median. He is asked to state two values (the .33 and .67 fractiles) such that the entire range is divided into three equally likely sections.

The most common calibration analysis is to calculate the interquartile index, which is the percent of items for which the true value falls inside the interquartile range (i.e., larger than the value associated with the 25th fractile, but smaller than the value associated with the 75th fractile), and to calculate the "surprise index," which is the percent of true values that fall outside the most extreme fractiles assessed. The perfectly calibrated person will, in the long run, have an interquartile index of 50. When the most extreme fractiles assessed are .01 and .99, then the perfectly calibrated person will have a surprise index of 2.

The impetus for investigating the calibration of probability density functions came from an unpublished paper by Alpert and Raiffa (1969), surely the most referenced rough draft in the literature of decision making. Alpert and Raiffa worked with four groups of subjects, all students enrolled in courses given by the Harvard Business School, and all familiar with the fundamentals of decision analysis. In their first experiment, all subjects assessed five fractiles, three of which were .25, .50, and .75. The extreme fractiles were, however, different for the different subgroups, .01 and .99 (Group A); .001 and .999 (Group B); "the minimum possible value" and "the maximum possible value" (Group C); and "astonishingly low" and "astonishingly high" (Group D). The interquartile and surprise indices for these four subgroups are shown in Table 1. Alpert and Raiffa, discouraged by the enormous number of surprises, then ran three additional groups who, after assessing 10 uncertain quantities, received feedback in the form of an extended report and explanation of the results, along with perorations that in the future the subjects should "*Spread Those Extreme Fractiles!*" (p. 13). The subjects then responded to 10 new uncertain quantities. Results before and after training are shown in Table 1. All groups showed some improvement with training. The greatest changes were shown by Group 4, the only group of subjects who were <u>not</u> exclusively from the Harvard Business School, but were enrolled in a decision analysis course designed for students from other departments.

Alpert and Raiffa experimented with fitting a beta function to the .25, .50, and .75 fractiles for a few subjects' responses to proportion questions (e.g., what proportion of students answering this questionnaire prefer Bourbon to Scotch?). The extreme fractiles of the fitted beta, rather than those the subjects actually gave, were used to compute the surprise index. This technique led to no improvement, suggesting that

TABLE 1

Calibration Summary for Continuous Items:

Percent of True Values Falling Within Interquartile Range and Outside the Extreme Fractiles

	N[a]	Interquartile Index[b]	Surprise Index	
		Observed	Observed	Ideal
Alpert & Raiffa (1969)				
Group 1-A (.01, .99)	880	33	46	2
Group 1-B (.001, .999)	500		40	.2
Group 1-C ("min" & "max")	700		47	?
Group 1-D ("astonishingly high/low")	700		38	?
Groups 2 & 3 Before	1670	33	39	2
After	1670	44	23	2
Group 4 Before	600	36	21	2
After	600	43	9	2
Hession & McCarthy (1974)	2035	25	47	2
Selvidge (1975)				
Five Fractiles	400	56	10	2
Seven Fractiles (incl. .1 & .9)	520	50	7	2
Schaefer & Borcherding (1973)				
1st Day, Fractiles	396	23	39	2
4th Day, Fractiles	396	38	12	2
1st Day, Hypothetical Sample	396	16	50	2
4th Day, Hypothetical Sample	396	48	6	2
Pickhardt & Wallace (1974)				
Group 1, First Round	?	39	32	2
Fifth Round	?	49	20	2
Group 2, First Round	?	30	46	2
Sixth Round	?	45	24	2
Pratt & Pratt (Personal Communication)				
"Astonishingly high/low"	175	37	5	?
Brown (1973)	414	29	42	2
Seaver, von Winterfeldt, & Edwards (1975)				
Fractiles	160	42	34	2
Odds-Fractiles	160	53	24	2
Probabilities	180	57	5	2
Odds	180	47	5	2
Log Odds	140	31	20	2
Murphy & Winkler (1974)				
Extremes were .125 & .875	132	45	27	25
Murphy & Winkler (this volume)				
Extremes were .125 & .875	432	54	21	25
Staël von Holstein (1971)	1269	27	30	2

[a] N is the total number of assessed distributions.

[b] The ideal percent of events falling within the interquartile range is 50, for all experiments except Brown (1973). He elicited the .30 and .70 fractiles, so the ideal is 40%.

the problem does not reside solely in subjects' inability to give sufficiently extreme .01 and .99 fractiles, but in their .25 and .75 fractiles as well.

Hession and McCarthy (1974) collected data comparable to Alpert and Raiffa's first session, using 55 uncertain quantities and 37 graduate students as subjects. In their instructions, they urged subjects to make certain that the interval between the .25 fractile and the .75 fractile did indeed capture half of the probability. "Later discussion with individual subjects made it clear that this consistency check resulted in most cases in a readjustment, decreasing the interquartile range originally assessed" (p. 7), thus making matters worse! This instructional emphasis, not used by Alpert and Raiffa, may explain why Hession and McCarthy's subjects were so badly calibrated, as shown in Table 1.

Hession and McCarthy also gave their subjects a number of "personality" tests they thought might be related to individual differences in calibration: the F (Authoritarian) Scale, the Dogmatism Scale, the Gough-Sanford Rigidity Scale, Pettigrew's Category-width Scale, and a group-administered intelligence scale. The correlations of these tests with the interquartile index and the surprise index across subjects were mostly quite low, although the F scale showed a hint of a relationship with calibration, correlating -.31 with the interquartile score and +.47 with the surprise score (N = 28).

Selvidge (1975) extended Alpert and Raiffa's work by first asking subjects four questions about themselves (e.g., do you prefer Scotch or Bourbon?). The responses were then used to find the true answer for what we will call "group-generated" uncertain quantities (e.g., how many of the 500 students answering the questionnaire preferred Scotch to Bourbon?). One group gave five fractiles, .01, .25, .5, .75, and .99. Another group gave

those five plus two others, .1 and .9. As shown in Table 1, the group with two additional fractiles did better. These results are not as different from the results of Alpert and Raiffa as they appear. Two of Alpert and Raiffa's uncertain quantities were group-generated proportions which were similar to Selvidge's items. On these two items only, Alpert and Raiffa found 58% in the interquartile range and 17% surprises. These results are much more similar to Selvidge's results than were their results for the entire 10-item set. Selvidge also reported surprise indices of 10% for extremes of .01 and .99 and 24% for extremes of .1 and .9, using five fractiles. Finally, when she asked subjects to give .25, .5 and .75 first, and then to give .01 and .99, she got fewer surprises (8%) than when the order was reversed (16%).

Schaefer and Borcherding (1973) explored the effects of training. They ran 22 university student subjects for four sessions, using 18 group-generated proportions per session. Each subject used two assessment techniques: (1) the fractile method (.01, .125, .25, .5, .75, .875, .99), and (2) the hypothetical sample method. In the latter method, subjects are asked to state the sample size, n, and the number of successes, r, of a hypothetical sample which best reflects their knowledge about the uncertain quantity. The larger n is, the more certain they are of the true value of the proportion. The ratio r/n reflects the mean of the distribution of their uncertainty. Subjects had great difficulty with this method, despite instructions which included examples of beta distributions. After every session subjects were given extensive feedback, with emphasis on their own and the groups' calibration. The results from the first and last sessions are shown in Table 1. Improvement was found for both methods. Results from the hypothetical sample method started out worse (50% surprises and only 16% in the interquartile range) but ended up better (6% surprises and 48%

in the interquartile range) than the fractile method.

Pickhardt and Wallace (1974) replicated Alpert and Raiffa's findings, with variations. Across several groups they reported 38 to 48% surprises before feedback, and not less than 30% surprises after feedback. Two variations, using or not using course grade credit as a reward, and using or not using scoring rule feedback, made no difference in the number of surprises. Pickhardt and Wallace also studied the effects of extended training. Two groups of 18 and 30 subjects (number of uncertain quantities not reported) responded for five and six sessions with calibration feedback after every session. Modest improvement was found, as shown in Table 1.

Finally, Pickhardt and Wallace studied the effects of increasing knowledge on calibration in the context of a realistic decision-making exercise: a production simulation game called PROSIM. Thirty-two graduate students each made 51 assessments during a simulated 17 "days" of production scheduling. Each assessment concerned an event that would occur 1, 2 or 3 "days" hence. The closer the time of assessment to the time of the event, the more the subject knew about the event. This increased information did affect calibration: there were 32% surprises with 3-day lags, 24% with 2-day lags, and 7% with 1-day lags. No improvement was observed over the 17 "days" of the simulation.

Pratt[8] asked a single expert to predict movie attendance for 175 movies or double features shown in two local theaters over a period of more than one year. The expert assessed the median, quartiles, and "astonishingly high" and "astonishingly low" values. As shown in Table 1, the interquartile range tended to be too small. Despite the fact that the expert received outcome feedback throughout the experiment, the only evidence of improvement in calibration over time came in the first few days.

---

[8] J. W. Pratt, personal communication, October, 1975.

Brown (1973) reported calibration results for 31 subjects responding to 14 uncertain quantities with fractiles .01, .10, .30, .50, .70, .90, and .99. The results, shown in Table 1, are particularly discouraging, because each question was accompanied by extensive historical data (e.g., for "Where will the consumer price index stand in December, 1970?", subjects were given the consumer price index for every quarter between March, 1962, and June, 1970). For 11 of the questions, had the subjects given the historical minimum as their .01 fractile and the historical maximum as their .99 fractile, they would have had no surprises at all. The other three questions showed strictly increasing or strictly decreasing histories, and the true value was close to any simple approximation of the historical trend. The subjects must have been putting a large emphasis on their own erroneous knowledge to have given distributions so tight as to produce 42% surprises.

Brown also reported unpublished data of Norman Dalkey and Bernice Brown, who elicited quartile assessments for uncertain quantities and found, for 1,218 cases, 31% of the true answers fell inside the interquartile range.

Seaver, von Winterfeldt, and Edwards (1975) studied the effects of five different response modes on calibration. Two groups used the fractile method, responding in units of the uncertain quantity to either fractile (.01, .25, .50, .75, .99) or the odds equivalents of those fractiles (1:99, 1:3, 1:1, 3:1, 99:1). Three other groups responded with probabilities, odds, or odds on a log-odds scale to one-alternative questions which specified a particular value of the uncertain quantity (e.g., what is the probability that the population of Canada in 1973 exceeded 25 million?). Five such questions were given for each uncertain quantity. For each group, seven to nine subjects, undergraduate and graduate students, responded to 20 uncertain quantities. As shown in Table 1, the groups giving probabilistic and odds responses had distinctly better surprise indices than those using the fractile

method. The log odds response mode did not work out well.

Four experiments used weather forecasters for subjects. In two experiments Murphy and Winkler (1974; and in press), using the variable-width, fixed-probability parallel to the earlier described fixed-width, variable-probability experiment (which we analyzed as a discrete task), asked subjects to give five fractiles (.125, .25, .5, .75, .875) for tomorrow's high temperature. The results, shown in Table 1, indicate excellent calibration. These subjects had fewer surprises in the extreme 25% of the distribution than did most of Alpert and Raiffa's subjects in the extreme 2%! Murphy and Winkler found that the five subjects in the two experiments who used the variable-width technique were better calibrated than the four subjects using the fixed-width technique. Pitz (1974), however, using a within-subject design with 44 college-student subjects, reported that the fractile technique led to worse calibration than the fixed-width technique, as did Seaver, von Winterfeldt and Edwards (1975).

Peterson, Snapper and Murphy (1972) asked for only three fractiles (.25, .5, and .75) for tomorrow's high temperature. Of 55 events, 51% fell inside the interquartile range, 16% fell on one of the boundaries, and 33% fell outside. This bit of data contains no evidence of poor calibration.

Staël von Holstein (1971) used three fixed-interval tasks: Average temperature tomorrow and the next day (dividing the entire response range into 8 categories), average temperature four and five days from now (8 categories), and total amount of rain in the next five days (4 categories). From each set of responses (4 or 8 probabilities summing to 1.0), he estimated the underlying cumulative density function. He then combined the 1,269 functions given by 28 participants. He reported an undue number

of surprises: 25% of the true answers fell below the inferred .07 fractile, and 25% fell above the .79 fractile. Using the group cumulative density function shown in his paper, we have estimated the surprise and interquartile indices (see Table 1). In contrast to the studies by Murphy and Winkler and by Peterson, Snapper and Murphy, these weather forecasters were quite poorly calibrated. Staël von Holstein's task was essentially similar to Murphy and Winkler's (1974) fixed-interval task. We have reviewed the former here and the latter in the section on discrete tasks simply because that is the way the authors summarized their data.

Barclay and Peterson (1973) compared the tertile method (i.e., the fractiles .33 and .67) with a "point" method in which the assessor is asked to give the modal value of the uncertain quantity, and then two values, one above and one below the mode, each of which is half as likely to occur as is the modal value (i.e., points for which the probability density function is half as high as at the mode). Using 10 almanac questions as uncertain quantities and 70 students at the Defense Intelligence School in a within-subject design, they found for the tertile method that 29% (rather than 33%) of the true answers fell in the central interval. For the point method, only 39% fell between the two half-probable points, whereas, for most distributions, approximately 75% of the density falls between these points.

Pitz (1974) reported several results using the tertile method. For 19 subjects estimating the populations of 23 countries, he found only 16% of the true values falling inside the central 33 percentile. He called this effect "hyperprecision." In another experiment he varied the items according to the depth and richness of knowledge he presumed his subjects to have. With populations of countries (low knowledge) he found 23% of

the true values in the central third; with heights of well-known buildings (middling knowledge), 27%; and with ages of famous people (high knowledge), 47%, the last being well above the expected 33%. In yet another study, he asked six subjects to assess tertiles, and a few days later to choose among bets based on their own tertile values. He found a strong preference for bets involving the central region, just the reverse of what their too-tight intervals should lead them to. Pitz suggested that the point estimate (the most likely value of the quantity) was over-controlling their choices.

The overwhelming evidence from research on uncertain quantities is that people's probability distributions tend to be too tight. The assessment of extreme fractiles is particularly prone to bias. Training improves calibration somewhat. Experts sometimes perform well (Murphy and Winkler, 1974, in press; Peterson, et al., 1972), sometimes not (Staël von Holstein, 1971). There is only scattered evidence that difficulty is related to calibration for continuous propositions. Pitz (1974) found such an effect, and Pickhardt and Wallace's (1974) finding that 1-day lags led to fewer surprises than 3-day lags in their simulation game is relevant here. Several studies (e.g., Barclay and Peterson, 1973; Murphy and Winkler, 1974) have reported a correlation between the spread of the assessed distribution and the absolute difference between the assessed median and the true answer, indicating that subjects do have a partial sensitivity to how much they do or do not know. This finding parallels the finding, with discrete propositions, of a correlation between percent correct and mean response. Pratt's expert showed no such correlation.[9]

DISCUSSION

Why should an assessor worry about being well calibrated? Von Winterfeldt and Edwards (1973) have shown that, in most real-world decision problems,

[9] J. W. Pratt, personal communication, November 13, 1975.

fairly large errors make little difference in the expected gain; "A suboptimal choice does not seriously hurt the decision maker as long as the alternative selected is not grossly away from the optimum" (p. 1). We can see at least two types of situations in which calibration does make a difference. First, in a two-alternative situation, the payoff function can be quite steep in the crucial region. Suppose your doctor must decide the probability that you have condition A, and should receive treatment A, versus having condition B and receiving treatment B. Suppose that the utilities are such that treatment A is better if the probability that you have condition A is $\geq .4$, as shown in Figure 15. If the doctor assesses the probability that you have A as $p(A)=.45$, but is poorly calibrated, so that he should have said .35, then he would treat you for B instead of A and you would lose quite a chunk of expected utility. Real-life utility functions of just this type are shown in Fryback (1974).

Secondly, even if the expected loss function for poor calibration is quite flat, the payoffs may be so large, and the errors so large, that the expected loss looms large. Weatherwax (1975), in critiquing the $3 million Rasmussen report on nuclear power safety (AEC, 1974) noted that "at each level of the analysis a log-normal distribution of failure rate data was assumed with 5 and 95 percentile limits defined" (p. 31). The research reviewed here suggests that distributions built from assessments of the .05 and .95 fractiles may be grossly biased. If such assessments are made at several levels of an analysis, with each assessed distribution being too narrow, the errors will not cancel each other, but will compound. And because the costs of nuclear disasters are large, the expected loss from such errors could be enormous.

If proper calibration is important, how can it be achieved? One way is to externally recalibrate the assessments people make. External

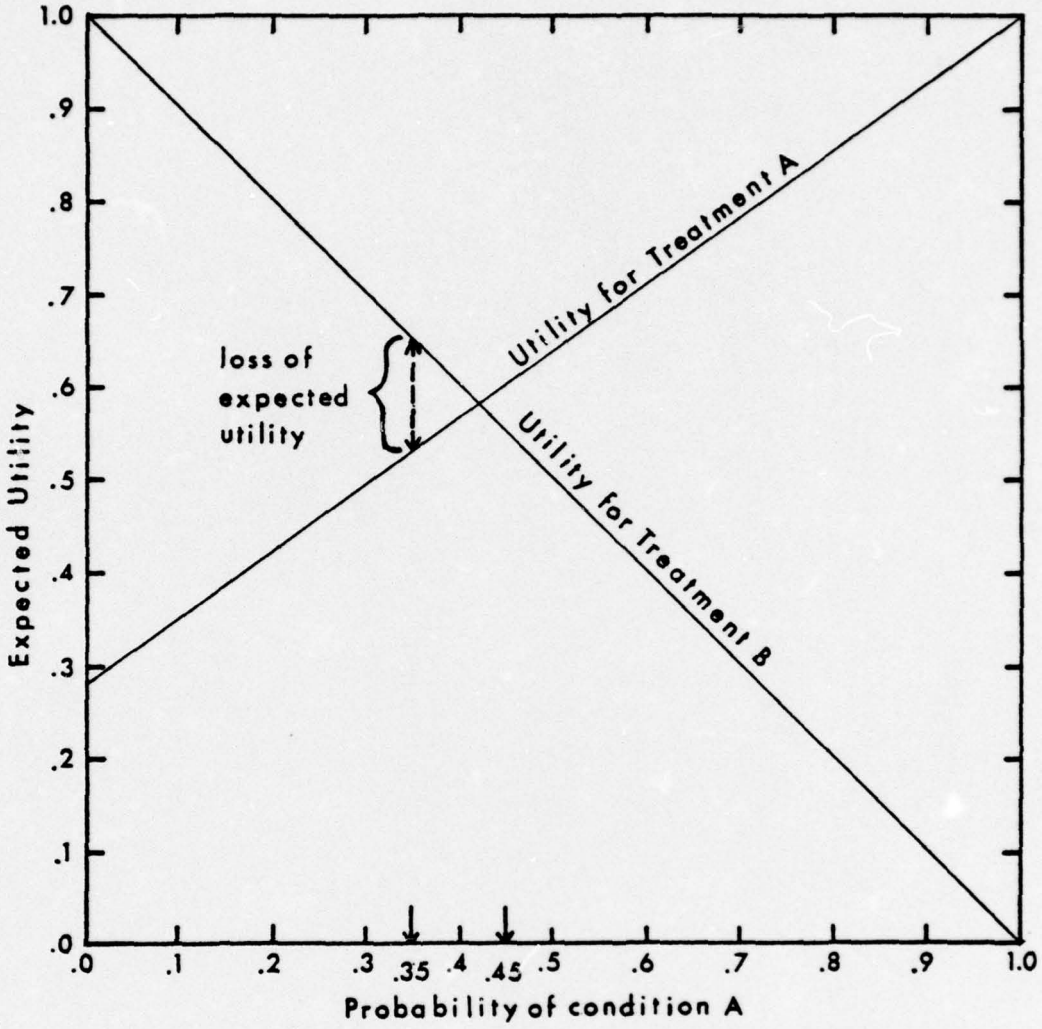


Figure 15

Loss of Utility due to Bad Calibration in a Medical Example

recalibration consists of collecting a set of assessments for items with known answers, fitting a model to the data, and substituting, in future assessments, the response predicted from the model for the response given by the assessor. The technical difficulties confronting recalibration are substantial. When eliciting the assessments to be modeled, one would have to be careful not to give the assessors any more feedback than they normally receive, for fear of their changing their calibration as it is being measured. As Savage (1971) pointed out, ". . . you might discover with experience that your expert is optimistic or pessimistic in some respect and therefore temper his judgments. Should he suspect you of this, however, you and he may well be on the escalator to prediction " (p. 796). One would also have to be quite confident that the real world matches, in difficulty, the known world on which their calibration is measured.

The theoretical objections to external recalibration may be even more serious than the practical objections. An assessor who consistently follows the axioms of probability theory can still be badly calibrated. The numbers produced by a recalibration process on such an assessor will not, in general, follow those axioms (for example, the numbers associated with mutually exclusive and exhaustive events will not always sum to one, nor will it be generally true that $P(A) \cdot P(B) = P(A,B)$ for independent events); hence, these new numbers cannot be called probabilities.

A more fruitful approach would be to train assessors to become well calibrated. The literature reviewed here gives us modest optimism that training might be successful. Yet we believe that the development of efficient training methods depends on our understanding of what is going on in a person's head when probabilities are assessed; this understanding depends on the development of good psychological theory.

The most striking aspect of the literature reviewed here is its "dust-bowl empiricism." Psychological theory is largely absent, either as motivation for the research or as explanation of the results. Much of the research seems motivated by simple questions beginning "What would happen if we. . . ?". Much of the interest in the research is in its potential applications. If people are going to have to assess probabilities in the course of making important future decisions, let us figure out the best way to do it. We can not help feeling that a better understanding of the psychological underpinnings of these findings would speed the solution to these applied problems.

Not all authors have avoided theorizing. Tversky and Kahneman (1974) and Slovic (1972) believe that, as a result of limited information-processing abilities, people adopt simplifying rules or heuristics. Although generally quite useful, these heuristics can lead to severe and systematic errors. For example, the tendency of people to give unduly tight distributions when assessing uncertain quantities could reflect the heuristic called "anchoring and adjustment." When asked about an uncertain quantity, one naturally thinks first of a point estimate, the most likely value. This value then serves as an anchor. To give the 25th or 75th percentile, one must adjust this anchor downwards or upwards. But the anchor has such a dominating influence that the adjustment is insufficient; hence the fractiles are too close together, yielding overconfidence. When, however, the experimenter provides a value, and the subject must supply a probability, the natural anchor is the first probability one thinks of. If that first probability thought of is .5 (reflecting initial uncertainty about whether the true value is above or below the value provided), then insufficient adjustment from this natural anchor will result in underconfidence. Tversky and Kahneman report data supporting this view. Pitz's (1974) data

in Figure 11, however, show overconfidence when a single value of the uncertain quantity is given to the subject. If these subjects were using the anchoring and adjustment heuristic, .5 was not their anchor.

Pitz (1974), too, believes that people's information-processing capacity and working memory capacity are limited. He suggests that people set up complex problems serially, working through a portion at a time. To reduce cognitive strain, people ignore the uncertainty in their solutions to the early portions of the problem in order to reduce the complexity of the calculations in later portions. This could lead to too-tight distributions and overconfidence. Pitz also suggests that one way people estimate their own uncertainty is by seeing how many different ways they can arrive at an answer, that is, how many different serial solutions they can construct. If many are found, people will recognize their own uncertainty; if few are found, they will not. The richer the knowledge base from which to build alternative structures, the less the tendency towards overconfidence. This was the reasoning that led Pitz to gather the data of Figure 11, which support his hypothesis.

These considerations are not full-fledged theories, but they may help us to gain understanding of how people think probabilistically. Another notion that may be helpful is coding. How do we code in our minds the outcomes we receive? Surely not the way we have coded, on paper, the data needed to plot a calibration curve.

> A person could conceivably learn whether his judgments are externally calibrated by keeping a tally of the proportion of events that actually occur among those to which he assigns the same probability. However, it is not natural to group events by their judged probability. In the

> absence of such grouping it is impossible for an individual to discover, for example, that only 50 percent of the predictions to which he has assigned a probability of .9 or higher actually came true. (Tversky & Kahneman, 1974, p. 1130)

In addition, as Fischhoff and Beyth (1975) found, even when subjects were forced to assess probabilities, they later altered their memory of these probabilities. Specifically, they remembered assigning higher probabilities than they actually had to events which later occurred and lower probabilities than they had to events which did not occur. To the extent that we do code events by probabilitic categories, we bias our coding towards overconfidence. "The judge who is insufficiently aware of the surprises the past held for him, and of the need to improve his performance, seems likely to continue being surprised by what happens in the future" (Fischhoff & Beyth, 1975, p. 15).

In conclusion, it seems appropriate to summarize what we know about calibration. We may characterize our knowledge as falling into one of three states: understanding, confusion, ignorance.

Understanding reigns when we have extensive evidence pointing at a common conclusion which any theory must accommodate. Understandings are, as might be expected, fairly scarce. One is that, as a result of subjects' failure to discriminate different levels of uncertainty adequately, different calibration curves emerge for tests with different levels of difficulty. A second conclusion is that the most common form of miscalibration is overconfidence. Nearly all the data about uncertain quantities point in this direction, as do the discrete-proposition data for all but the easiest tasks. If overconfidence is further evidence of a general tendency toward what Dawes (1976) calls "cognitive conceit," it

is crucial to understand its origins, limits and remedies. A third and more optimistic conclusion is that calibration can be somewhat improved by training.

Confusion reigns when studies of a given question point in contrary directions or when we must put our faith in a single study using but one of the many possible variations of experimental procedure and stimuli. Consider for example the symmetry or asymmetry of the curves in different full-range studies, or the contrary contrasts of the variable-width and fixed-width methods of Pitz (1974) and Murphy and Winkler (1975), or Hazard and Peterson's (1973) lonely finding that odds and probability judgments have similar calibration curves.

One partial solution to the problem of divergent findings is to increase our understanding of the sampling properties of calibration curves. Some conflicting results may be attributable to sampling variations. The second general solution (aside from collecting more data) is to improve our theoretical conceptualization of probability assessment tasks and of the factors which influence performance. Apparently divergent findings may be explained by previously unnoted differences in task characteristics such as difficulty level, instructions, or implicit loss functions.

When ignorance reigns, it is the job of any theory to advance interesting hypotheses and identify crucial issues. Even in lieu of developed theories, it is still possible to raise many questions that bear answering. What are the effects of varying instructions, e.g., ardently discouraging the use of .00 and 1.00? Are there any response modes particularly conducive to calibrated judgments? Should one restrict assessors to some fixed number of possible probability responses (say,

.5, .75, and .99) which reflects the number of meaningful discriminations that they can make? What is the effect of the number of alternatives on calibration? Are there individual differences in calibration and, if so, what distinguishes well-calibrated judges? Holding task difficulty constant, neither brains nor expertise appears to make much difference. We have recently found that with a half-range, two-alternative task, heavy reliance on the responses .50 and 1.00 (which might reflect lack of effort or perceived inability to make finer distinctions) is <u>not</u> a sign of inferior calibration. Other than task difficulty, what does make a difference? Even without theoretical advances, we have some work to do before reaching the bottom of empiricism's dust-bowl.

REFERENCES

Adams, J. K. A confidence scale defined in terms of expected percentages. American Journal of Psychology, 1957, 70, 432-436.

Adams, J. K. & Adams, P. A. Realism of confidence judgments. Psychological Review, 1961, 68, 33-45.

Adams, P. A. & Adams, J. K. Training in confidence judgments. American Journal of Psychology, 1958, 71, 747-751.

Alpert, M. & Raiffa, H. A progress report on the training of probability assessors. Unpublished manuscript, 1969.

Atomic Energy Commission. Reactor Safety Study. An assessment of accident risks in U. S. commercial power plants, WASH-1400 Draft. Washington, D.C.: The Commission, 1974.

Barclay, S. & Peterson, C. R. Two methods for assessing probability distribution. Decisions and Designs, Inc. Technical Report, 73-1, 1973.

Brier, G. W. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 1950, 75, 1-3.

Brown, T. A. An experiment in probabilistic forecasting. RAND Report R-944-ARPA, 1973.

Brown, T. A. & Shuford, E. H. Quantifying uncertainty into numerical probabilities for the reporting of intelligence. RAND Report R-1185-ARPA, 1973.

Clarke, F. R. Confidence ratings, second-choice responses, and confusion matrices in intelligibility tests. Journal of the Acoustical Society of America, 1960, 32, 35-46.

Cooke, W. E. Forecasts and verifications in Western Australia. Monthly Weather Review, 1906a, 34, 23-24.

Cooke, W. E. Weighting forecasts. Monthly Weather Review, 1906b, 34, 274-275.

Dawes, R. M. Shallow psychology. In J. S. Carroll and J. W. Payne (Eds.), Cognition and Social Behavior. Potomac, Md.: Lawrence Erlbaum Associates, 1976.

Decker, L. R. & Pollack, I. Confidence ratings and message reception for filtered speech. Journal of the Accoustical Society of America, 1958, 30, 432-434.

de Finetti, B. Foresight: Its logical laws, its subjective sources. Annales de l'Institut Henri Poincaré, 1937, 7, reprinted in English; Studies in subjective probability, H. E. Kyburg, Jr. & H. E. Smokler (Eds.), New York: Wiley, 1964.

Fischhoff, B. & Beyth, R. "I knew it would happen"--remembered probabilities of once-future things. Organizational Behavior and Human Performance, 1975, 13, 1-16.

Fischhoff, B. & Lichtenstein, S. Calibration potpourri, manuscript in preparation.

Fryback, D. G. Use of radiologists' subjective probability estimates in a medical decision making problem, Michigan Mathematical Psychology Program, Report 74-14, 1974, University of Michigan, Ann Arbor, Michigan.

Green, D. M. & Swets, J. A. Signal Detection Theory and Psychophysics. New York: Wiley, 1966.

Hazard, T. H. & Peterson, C. R. Odds versus probabilities for categorical events. Decision & Designs, Inc. Technical Report, 73-2, 1973.

Hession, E. & McCarthy, E. Human performance in assessing subjective probability distribution. Unpublished manuscript, September, 1974.

Lichtenstein, S. & Fischhoff, B. Do those who know more also know more about how much they know?, Oregon Research Institute Research Bulletin, 1976, 16, 1.

Murphy, A. H. Scalar and vector partitions of the probability score: Part I. Two-state situation. Journal of Applied Meteorology, 1972, 11, 273-282.

Murphy, A. H. A new vector partition of the probability score. Journal of Applied Meteorology, 1973, 12, 595-600.

Murphy, A. H. A sample skill score for probability forecasts. Monthly Weather Review, 1974, 102, 48-55.

Murphy, A. H. & Winkler, R. L. Forecasters and probability forecasts: Some current problems. Bulletin of the American Meteorological Society, 1971, 52, 239-247.

Murphy, A. H. & Winkler, R. L. Subjective probability forecasting experiments in meteorology: Some preliminary results. Bulletin of the American Meteorological Society, 1974, 55, 1206-1216.

Murphy, A. H. & Winkler, R. L. The use of credible intervals in temperature forecasting: Some experimental results. [This volume.]

Oskamp, S. The relationship of clinical experience and training methods to several criteria of clinical prediction. Psychological Monographs, 1962, 76, 28 (Whole No. 547).

Peterson, C. R., Snapper, K. J. & Murphy, A. H. Credible interval temperature forecasts. Bulletin of the American Meteorological Society, 1972, 53, 966-970.

Phillips, L. D. & Lichtenstein, S. Modeling the calibration of probability assessments, in preparation.

Phillips, L. D. & Wright, [This volume.]

Pickhardt, R. C. & Wallace, J. B. A study of the performance of subjective probability assessors. Decision Sciences, 1974, 5, 347-363.

Pitz, G. F. Subjective probability distributions for imperfectly known quantities. Knowledge and Cognition, L. W. Gregg (Ed.), New York: Wiley, 1974, 29-41.

Pollack, I. & Decker, L. R. Confidence ratings, message receptions, and the receiver operating characteristic. Journal of the Acoustical Society of America, 1958, 30, 286-292.

Root, H. E. Probability statements in weather forecasting. Journal of Applied Meteorology, 1962, 1, 163-168.

Sanders, F. The evaluation of subjective probability forecasts. Cambridge, Mass.: Institute of Technology, Department of Meteorology. Scientific Report #5, 1958.

Savage, L. J. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 1971, 66, 336, 783-801.

Schaefer, R. E. & Borcherding, K. The assessment of subjective probability distribution: A training experiment. Acta Psychologica, 1973, 37, 117-129.

Schlaifer, R. Computer programs for elementary decision analysis. Boston: Harvard University Press, 1971.

Seaver, D., von Winterfeldt, D. & Edwards, W. Eliciting subjective probability distributions on continuous variables. University of Southern California, Social Science Research Institute, Technical Report 75-8, 1975.

Selvidge, J. Experimental comparison of different methods for assessing the extremes of probability distributions by the fractile method. Management Science Report Series, Report 75-13, 1975, Graduate School of Business Administration, University of Colorado, Boulder, Colorado.

Shuford, E. & Brown, T. A. Elicitation of personal probabilities and their assessment. Instructional Science, 1975, 4, 137-188.

Sieber, J. E. Effects of decision importance on ability to generate warranted subjective uncertainty. Journal of Personality and Social Psychology, 1974, 30, 686-694.

Slovic, P. From Shakespeare to Simon: Speculations--and some evidence--about man's ability to process information. Oregon Research Institute Research Bulletin, 1972, 12, 2.

Staël von Holstein, C.-A. S. An experiment in probabilistic weather forecasting, Journal of Applied Meteorology, 1971, 10, 635-645.

Staël von Holstein, C.-A. S. Probabilistic forecasting: An experiment related to the stock market. Organizational Behavior and Human Performance, 1972, 8, 139-158.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. Decision processes in perception. Psychological Review, 1961, 68, 301-340.

Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 1974, 185, 1124-1131.

United States Weather Bureau. Report on weather bureau forecast performance 1967-8 and comparison with previous years, Technical Memorandum WBTM FCST, 11, Office of Meteorological Operations, Weather Analysis and Prediction Division, Silver Spring, Md., March, 1969.

von Winterfeldt, D. & Edwards, W. Flat maxima in linear optimization models. Technical Report #011313-4-T, Engineering Psychology Laboratory, 1973, University of Michigan, Ann Arbor, Michigan.

Weatherwax, R. K. Virtues and limitations of risk analysis. Bulletin of the Atomic Scientists, 1975, 31, 29-32.

Williams, P. The use of confidence factors in forecasting. Bulletin of the American Meteorological Society, 1951, 32, 8, 279-281.

Winkler, R. L. & Murphy, A. H. "Good" probability assessors. Journal of Applied Meteorology, 1968a, 7, 751-758.

Winkler, R. L. & Murphy, A. H. Evaluation of subjective precipitation probability forecasts. Proceedings of the First National Conference on Statistical Meteorology, Hartford, Conn., May 27-29, 1968b, 148-157, American Meteorological Society, Boston, Massachusetts.

Research Distribution List

Department of Defense

Assistant Director (Environment and Life Sciences)
Office of the Deputy Director of Defense *Research and Engineering* (Research and Advanced Technology)
Attention: Lt. Col. Henry L. Taylor
The Pentagon, Room 3D129
Washington, DC 20301

Office of the Assistant Secretary of Defense (Intelligence)
Attention: CDR Richard Schlaff
The Pentagon, Room 3E279
Washington, DC 20301

Director, Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Program Management Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209
(two copies)

Administrator, Defense Documentation Center
Attention: DDC-TC
Cameron Station
Alexandria, VA 22314
(12 copies)

Department of the Navy

Office of the Chief of Naval Operations (OP-987)
Attention: Dr. Robert G. Smith
Washington, DC 20350

Director, Engineering Psychology Programs (Code 455)
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217
(three copies)

Assistant Chief for Technology (Code 200)
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 230)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Analysis Programs (Code 431)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Operations Research Programs (Code 434)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 436)
Attention: Dr. Bruce McDonald
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Information Systems Program (Code 437)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research (ONR)
International Programs (Code 1021P)
800 North Quincy Street
Arlington, VA 22217

Director, ONR Branch Office
Attention: Dr. Charles Davis
536 South Clark Street
Chicago, IL 60605

Director, ONR Branch Office
Attention: Dr. J. Lester
495 Summer Street
Boston, MA 02210

Director, ONR Branch Office
Attention: Dr. E. Gloye and Mr. R. Lawson
1030 East Green Street
Pasadena, CA 91106
(two copies)

Dr. M. Bertin
Office of Naval Research
Scientific Liaison Group
American Embassy – Room A-407
APO San Francisco 96503

Director, Naval Research Laboratory
Technical Information Division (Code 2627)
Washington, DC 20375
(six copies)

Director, Naval Research Laboratory (Code 2029)
Washington, DC 20375
(six copies)

Scientific Advisor
Office of the Deputy Chief of Staff
for Research, Development and Studies
Headquarters, U.S. Marine Corps
Arlington Annex, Columbia Pike
Arlington, VA 20380

Headquarters, Naval Material Command (Code 0331)
Attention: Dr. Heber G. Moore
Washington, DC 20360

Headquarters, Naval Material Command (Code 0344)
Attention: Mr. Arnold Rubinstein
Washington, DC 20360

Naval Medical Research and Development Command (Code 44)
Naval Medical Center
Attention: CDR Paul Nelson
Bethesda, MD 20014

Head, Human Factors Division
Naval Electronics Laboratory Center
Attention: Mr. Richard Coburn
San Diego, CA 92152

Dean of Research Administration
Naval Postgraduate School
Monterey, CA 93940

Naval Personnel Research and Development Center
Management Support Department (Code 210)
San Diego, CA 92152

Naval Personnel Research and Development Center (Code 305)
Attention: Dr. Charles Gettys
San Diego, CA 92152

Dr. Fred Muckler
Manned Systems Design, Code 311
Navy Personnel Research and Development Center
San Diego, CA 92152

Human Factors Department (Code N215)
Naval Training Equipment Center
Orlando, FL 32813

Training Analysis and Evaluation Group
Naval Training Equipment Center
(Code N-00T)
Attention: Dr. Alfred F. Smode
Orlando, FL 32813

Department of the Army

Technical Director, U.S. Army Institute for the Behavioral and Social Sciences
Attention: Dr. J.E. Uhlaner
1300 Wilson Boulevard
Arlington, VA 22209

Director, Individual Training and Performance Research Laboratory
U.S. Army Institute for the Behavioral and
and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Director, Organization and Systems Research Laboratory
U.S. Army Institute for the Behavioral and
Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Department of the Air Force

Air Force Office of Scientific Research
Life Sciences Directorate
Building 410, Bolling AFB
Washington, DC 20332

Robert G. Gough, Major, USAF
Associate Professor
Department of Economics, Geography and
Management
USAF Academy, CO 80840

Chief, Systems Effectiveness Branch
Human Engineering Division
Attention: Dr. Donald A. Topmiller
Wright-Patterson AFB, OH 45433

Aerospace Medical Division (Code RDH)
Attention: Lt. Col. John Courtright
Brooks AFB, TX 78235

Other Institutions

The Johns Hopkins University
Department of Psychology
Attention: Dr. Alphonse Chapanis
Charles and 34th Streets
Baltimore, MD 21218

Institute for Defense Analyses
Attention: Dr. Jesse Orlansky
400 Army Navy Drive
Arlington, VA 22202

Director, Social Science Research Institute
University of Southern California
Attention: Dr. Ward Edwards
Los Angeles, CA 90007

Perceptronics, Incorporated
Attention: Dr. Amos Freedy
6271 Variel Avenue
Woodland Hills, CA 91364

Director, Human Factors Wing
Defense and Civil Institute of
Environmental Medicine
P.O. Box 2000
Downsville, Toronto
Ontario, Canada

Stanford University
Attention: Dr. R.A. Howard
Stanford, CA 94305

Montgomery College
Department of Psychology
Attention: Dr. Victor Fields
Rockville, MD 20850

General Research Corporation
Attention: Mr. George Pugh
7655 Old Springhouse Road
McLean, VA 22101

Oceanautics, Incorporated
Attention: Dr. W.S. Vaughan
3308 Dodge Park Road
Landover, MD 20785

Director, Applied Psychology Unit
Medical Research Council
Attention: Dr. A.D. Baddeley
15 Chaucer Road
Cambridge, CB 2EF
England

Department of Psychology
Catholic University
Attention: Dr. Bruce M. Ross
Washington, DC 20017

Stanford Research Institute
Decision Analysis Group
Attention: Dr. Allan C. Miller III
Menlo Park, CA 94025

Human Factors Research, Incorporated
Santa Barbara Research Park
Attention: Dr. Robert R. Mackie
6780 Cortona Drive
Goleta, CA 93017

University of Washington
Department of Psychology
Attention: Dr. Lee Roy Beach
Seattle, WA 98195

Eclectech Associates, Incorporated
Post Office Box 179
Attention: Mr. Alan J. Pesch
North Stonington, CT 06359

Hebrew University
Department of Psychology
Attention: Dr. Amos Tversky
Jerusalem, Israel

Dr. T. Owen Jacobs
Post Office Box 3122
Ft. Leavenworth, KS 66027

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

Calibration of Probabilities: The State of the Art.

5. TYPE OF REPORT & PERIOD COVERED

Technical rept.

6. PERFORMING ORG. REPORT NUMBER

ORI Report No.: DDI-3

7. AUTHOR(s)

Sarah Lichtenstein
Baruch Fischhoff
L. D. Phillips

8. CONTRACT OR GRANT NUMBER(s)

Prime Contract No.: N00014-76-C-0074
Subcontract No.: 75-030-0712

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Oregon Research Institute
P.O. Box 3196
Eugene, Oregon 97403

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

ARPA Order-3052

11. CONTROLLING OFFICE NAME AND ADDRESS

Defense Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, VA 22209

12. REPORT DATE

August 1976

13. NUMBER OF PAGES

73

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

Office of Naval Research
800 North Quincy Street
Arlington, VA 222]7

15. SECURITY CLASS. (of this report)

Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Support for this research performed by Oregon Research Institute was provided by the Advanced Research Projects Agency of the Department of Defense and was monitored under Contract N00014-76-C-0074 with the Office of Naval Research, under subcontract from Decisions and Designs, Inc.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Calibration
Probability Assessment
Uncertain Quantities

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An important criterion for evaluating probability assessors is their degree of calibration. A probability assessor is well calibrated if, over the long run, for all statements assigned a given probability (e.g., the probability is .65 that "Romania will maintain its current relation with People's China"), the proportion that is true is equal to the probability assigned. For example, if you are well calibrated, then across all the many occasions that you assign a probability of .8, in the long run 80% of them

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

390 664

should turn out to be true. If, instead, only 70% are true, you are not well calibrated, you are overconfident. If 95% of them are true, you are underconfident. In the last few years, there has developed an extensive literature about calibration, reporting both laboratory and real-world experiments. The present report reviews this literature, looking for findings that can be used to improve decisions. Among the major findings are the following:

1. Weather forecasters, who typically have had several years of experience in assessing probabilities, are well calibrated.

2. Other experiments, using a wide variety of tasks and subjects, show that people are generally quite poorly calibrated. In particular, people act as though they can make much finer distinctions in their degree of uncertainty than is actually the case.

3. Overconfidence is found in most tasks; that is, people tend to overestimate how much they know.

4. Despite the abundant evidence that untutored assessors are badly calibrated, there is little research showing how and how well these deficiencies can be overcome through training.